KORPUS-ADAPTIVE EIGENNAMENERKENNUNG

Dissertation
zur Erlangung der Würde eines
Doktors der Philosophie (Dr. phil.)

Vorgelegt im Fach Computerlinguistik
Fachbereich Geisteswissenschaften der
Universität Duisburg-Essen
von
Marc Rössler aus Winterthur (CH)

Gutachter der Arbeit:
Prof. Dr. Wolfgang Hoeppner, Prof. Dr. Katharina Morik, Prof. Dr. Jürgen Biehl

Datum der Disputation: 21.12.2006

Danksagung

Herzlich danken möchte ich meinem Betreuer Prof. Dr. Wolfgang Hoeppner. Sein Vertrauen, seine Offenheit gegenüber neuen Wegen und seine Unterstützung haben mein Dissertationsprojekt erst möglich gemacht. Meiner Zweitgutachterin Prof. Dr. Katharina Morik möchte ich für eine Vielzahl fruchtbarer Diskussionen, Anregungen und Erklärungen danken, die mir Zugang zur faszinierenden Welt des Maschinellen Lernens verschafft haben.

Bedanken möchte ich mich auch bei den vielen Menschen, deren Zeit, Unterstützung und Hilfestellung ich während dieser Zeit in Anspruch genommen habe. Dazu gehören unter anderem Claudia Basener, Prof. Dr. Jürgen Biehl, Nina Delvos, Sandra Kübler, HD Dr. Frank Pointner, Nino Simunic, die Teilnehmer des LS-8 Frühstück-Kolloquiums, Jorn Veenstra, Anne-Katharin Venohr, Lore Venohr, Andreas Wagner und Issai Zaks.

Nicht unerwähnt bleiben darf mein Sohn Yannik, der so verständnisvoll damit umging, dass die Gedanken seines Vaters ganz oft in weit entfernten Sphären herumkreisten und meine Freundin Irma, deren Liebe mir so viel Kraft gibt.

Danken möchte ich darüber hinaus meinen Eltern, denen ich nicht nur ein Studium, sondern noch viel mehr zu verdanken habe.

Duisburg, im Januar 2007

Für meine viel zu früh verstorbene Schwester Leni

Inhaltsverzeichnis

Abbildungsverzeichnis

Die Nummern der Abbildungen bestehen aus der Kapitelnummer und einer fortlaufenden Nummerierung.

1. Einleitung

Große Fortschritte in der Informationstechnologie ermöglichen den elektronischen Zugriff auf Informationen in einem kaum vorstellbaren Umfang. Ein Großteil dieser Informationen liegt in natürlichsprachlicher Form vor. Zwar ermöglichen Suchmaschinen den direkten Zugriff auf alle Textpassagen, die eine bestimmte Zeichenkette enthalten, doch sind darüber hinausgehende semantische Erschließungsverfahren wünschenswert. Der rein zeichenbasierte Zugriff vernachlässigt wichtige Dimensionen der Sprache und führt dazu, dass vorhandene Information nicht gefunden wird und damit für den Anwender nicht existiert oder aber gefundene Information nicht in einer Art und Weise vorliegt, die eine effiziente Weiterverarbeitung ermöglicht. Die Named Entity Recognition (NER) gehört zu den automatischen Verfahren zur semantischen Erschließung von Texten und dient - etwas vereinfacht ausgedrückt - der Erkennung aller in Texten vorkommenden Eigennamen, die einer im Vorfeld definierten Kategorie angehören. Kategorien von Eigennamen sind beispielsweise Namen von Personen, Organisationen, geographischen Objekten, aber auch fachspezifische Namen, etwa aus der Biomedizin. Das Ergebnis eines NER-Systems sind Texte mit markierten und kategorisierten Eigennamen. Die automatische Markierung dieser Einheiten erlaubt es, die Aufbereitung von Informationen effizienter und komfortabler zu gestalten und damit den Zugriff auf Informationen zu optimieren. Darüber hinaus ist die NER unverzichtbarer Bestandteil einer umfassenderen syntaktisch-semantischen Textanalyse.

In dieser Dissertation wird ein Korpus-adaptiver Ansatz zur NER vorgestellt, also ein NER-Verfahren, welches effizient auf neue Korpora bzw. Anwendungen angepasst werden kann. Um die Anforderungen an ein solches System und seine Einordnung in den aktuellen wissenschaftlichen Forschungsstand, aber auch den wissenschaftlichen Beitrag dieser Arbeit zu verdeutlichen, werden Grundlagen zu Named Entities, Named Entity Recognition und zum bisherigen Stand der Wissenschaft erarbeitet. Im Folgenden wird der Aufbau der Arbeit vorgestellt.

Kapitel 2 führt den Begriff der Named Entity (NE) ein und stellt dar, in welchen Bereichen eine klare Abgrenzung fehlt. In welcher Art und Weise sich diese Unschärfe bei der menschlichen Annotation auswirkt, also der Anwendung der Definitionen auf Texte, wird anhand dreier annotierter Korpora untersucht. Dabei wirft insbesondere die NE-Annotation deutscher Texte interessante Fragen zur Abgrenzung von NEs auf. Anschließend wird ein pragmatischer Umgang mit der vorhandenen Unschärfe des Begriffs vorgeschlagen, welcher

der anwendungsorientierten Perspektive gerecht wird, die dieser Arbeit und den meisten Ansätzen zur NER zugrunde liegt.

Kapitel 3 widmet sich der Aufgabe der Erkennung und Klassifikation von NEs, also der eigentlichen NER. Nach der Motivation der Aufgabenstellung und der Einführung der Evaluationsverfahren wird die Schwierigkeit der Aufgabe diskutiert und mögliche Lösungen aufgezeigt. Neben der Einführung der grundlegenden Begriffe der internen und der externen Evidenz werden verschiedene Arten der Generalisierung erörtert, welche als Grundlage zur Modellbildung dienen können. Hierbei wird zwischen der modell- und der datenorientierten Generalisierung unterschieden. Während der modellorientierte Ansatz mittels linguistisch anerkannter Methoden generalisiert, beschränkt sich der datenorientierte Ansatz auf die Eigenschaften, die sich direkt in den vorhandenen Daten beobachten lassen. Des Weiteren beschäftigt sich Kapitel 3 mit lexikalischen Ressourcen für die NER. Diese werden vorrangig unter dem Aspekt der semantischen Kategorisierung untersucht, wobei Fragen der gewählten Kategorien, aber auch der semantischen Mehrdeutigkeit behandelt werden. Da die NER nicht auf die Erkennung und Klassifikation einzelner Wörter beschränkt ist, sondern ganze Wortsequenzen erfassen muss, werden der Mehrwortcharakter von NEs und die damit verbundenen Schwierigkeiten diskutiert. Im letzten Abschnitt des Kapitels wird der Zusammenhang zwischen Texteinheiten und NEs dargestellt.

Im Kapitel 4 werden die bisherigen Arbeiten zur NER vorgestellt. Angestoßen durch die sechste MUC-Evaluationskampagne ([MUC-6 1995]) ist eine große Anzahl von Beiträgen zur NER entstanden. Unterteilt in regel- und lernbasierte Ansätze werden die wichtigsten dieser Arbeiten besprochen. Der Einsatz umfangreicher, nicht-annotierter Textsammlungen zum Erwerb von Wissen für die NER ist ein wichtiger Schwerpunkt der jüngeren Arbeiten zur NER und wird in einem eigenen Abschnitt besprochen. Die Auswertung bzw. Ausnutzung dieser meist nahezu umsonst vorhandenen Ressourcen ist äußerst attraktiv, da dieses Vorgehen die effizientere Entwicklung, aber auch Wartung und Anpassung von NER-Systemen verspricht. Der Einsatz nicht-annotierter Daten ist von zentraler Bedeutung für die Korpus-Adaptivität des eigenen Ansatzes. Der Begriff der Korpus-Adaptivität ist weitgehend unbekannt und wird deshalb eingeführt und definiert. Anhand der Diskussion existierender NER-Ansätze hinsichtlich ihrer Anpassungsfähigkeit an ein neues Korpus treten die Kriterien der Korpus-Adaptivität deutlich zutage.

Ausgehend von den Kriterien der Korpus-Adaptivität können wichtige Eckpunkte des in Kapitel 5 beschriebenen NER-Systems in Form von Entwurfsrichtlinien festgeschrieben

werden. Gemäß dieser Richtlinie ist ein lernbasiertes System erforderlich, welches auf einer durchgehend datenorientierten Generalisierung beruht und weitestgehend auf manuell erstellte Ressourcen verzichtet. Bei der Implementation wird die datenorientierte Generalisierung durch eine Kombination bewährter und neu entwickelter Verfahren erreicht, die unabhängig von linguistischen Werkzeugen und Methoden operieren. Durch die Beschränkung manuell erstellter Ressourcen auf ein annotiertes Korpus wird der Anteil dieser aufwändig zu erstellenden Wissensquellen sehr gering gehalten. Um die resultierende Wissensarmut des Ansatzes zu kompensieren, werden nicht-annotierte Daten ausgewertet. Diese liegen meist in großen Mengen vor und können mit geeigneten Methoden zur Erzeugung von Ressourcen eingesetzt werden. Der Einsatz der nicht annotierten Daten ist von der Idee angetrieben, dass ein einfacher Klassifizierer anhand der Auswertung seiner eigenen Klassifikationsentscheide verbessert werden kann. Dazu wird ein mit wenigen Ressourcen ausgestatteter Klassifizierer zur Erkennung von NEs in einem umfangreichen nicht-annotierten Korpus eingesetzt. Basierend auf den erkannten NEs wird versucht, weitere Hinweise zur Erkennung von NEs abzuleiten. Verbessern die gefundenen Hinweise die Erkennungsrate, so kann das Erkennen von NEs und die Ableitung weiterer Hinweise zur Erkennung von NEs wiederholt werden. Im Gegensatz zu anderen Ansätzen, welche nicht-annotierte Daten auswerten, werden hierbei keine Namenslisten oder zusätzliche Lernbeispiele extrahiert. Vielmehr werden die Lernbeispiele durch zusätzliche Merkmale angereichert. Dieses Vorgehen erst ermöglicht den Einsatz linearer Support Vektor Maschinen als Lernalgorithmus, da diese zwar sehr effizient mit hochdimensionalen Merkmalsräumen, nicht aber mit großen Mengen von Trainingsdaten umgehen können.

Die Evaluation dieses Ansatzes wird in Kapitel 6 beschrieben. Um eine detaillierte Auswertung zu erhalten, wird das entwickelte System in unterschiedlichen Konfigurationen eingesetzt und auf unterschiedliche Korpora angewandt. Anhand der Evaluationsergebnisse wird das System mit anderen Systemen verglichen und in einer abschließenden Diskussion bewertet. Die Arbeit endet mit einer Zusammenfassung und einem Ausblick.

2. Named Entities

Im Rahmen der Message Understanding Conferences wurde 1995 eine Teilaufgabe der Informationsextraktion definiert, die sog. Named Entity Recognition (NER). Der Begriff der Named Entity (NE) existierte vorher nicht und hat bisher auch keinen Eingang in die Sprachwissenschaft oder andere Disziplinen außerhalb der automatischen Sprachverarbeitung gefunden. Trotz der vielen Arbeiten zur NER existiert bis heute keine allgemeine und eindeutig anzuwendende Definition der NEs. Im Deutschen gibt es keine adäquate Übersetzung für NER; die manchmal verwendete „Eigennamenerkennung“ gibt den Begriff nur unzureichend wieder. Zwar kann die Aufgabe der NER verkürzt als Erkennung von Eigennamen beschrieben werden, doch erfordert die NER zusätzlich eine semantische Klassifizierung der erkannten Einheiten. Außerdem sind Eigennamen zwar zentral für die Definition von NEs, doch sind die beiden Begriffe nicht deckungsgleich: Einzig die Eigennamen der vorgegebenen Kategorien sind eine NE und nicht jede NE ist ein Eigenname. So ist etwa die Wortkette „*51. Internationale Kurzfilmtage Oberhausen*“ ein Eigenname, sie ist jedoch nur eine NE, wenn die semantische Klasse der Ereignisse oder Festivals als zu erkennend festgelegt ist. Auch kann eine NER-Aufgabe festlegen, dass „*Universität*“ als Koreferenz auf die „*Universität Duisburg-Essen*“ als NE der Klasse Organisation zu markieren ist, obwohl „*Universität*“ kein Eigenname ist.

Für zusätzliche Unklarheiten sorgt der für die NE so zentrale Begriff des Eigennamens. Obwohl dieser bereits in frühesten Arbeiten über Wortklassen auftaucht, so entzieht sich der Eigenname seit jeher einer vollständigen linguistischen Definition.

Ohne den Anspruch diese Unklarheiten vollständig zu beheben, wird in diesem Kapitel der Begriff der NE ausgeleuchtet. Dazu ist es erforderlich, einige Grundlagen zum Thema Eigennamen darzustellen. Darauf aufbauend werden anhand unterschiedlicher Anwendungen des NE Begriffs einige Schwierigkeiten und Grenzfälle aufgezeigt. Das Kapitel endet mit einer Zusammenfassung zur Unschärfe des NE Begriffs.

2.1. Über Eigennamen

Eine besondere Stellung im lexikalischen System wird den Eigennamen bereits in einer der ersten grammatischen Schriften zugewiesen. Dionysios Thrax (170-90 v.Ch.) unterscheidet in seiner Wortartenlehre die „vollwertigen Namen“, wie „Homer, Sokrates“, die eine

„individuelle Realität“ bezeichnen, von den Namen, die der „Bezeichnung dienen“, und „allgemein Seiendes“, wie „Mensch, Pferd“ benennen ([BAUER 1998: 33]).

Thrax’ Begriff vom „vollwertigen Namen“ entspricht der intuitiven Vorstellung vom Eigennamen, der zur Referenz auf genau ein konkretes Objekt der Welt dient und diesem ähnlich wie ein Etikett anhaftet. Das Phänomen der Eigennamen bzw. der Benennung ist seither aus vielerlei Perspektiven untersucht worden.

Frege führt den Eigennamen in die sprachphilosophische Diskussion ein ([FREGE 1892]), indem er nach der Semantik von Eigennamen fragt. Er geht davon aus, dass mit jedem Eigennamen zwei semantische Komponenten verbunden sind, nämlich erstens ein Referenzobjekt, welcher er als Bedeutung des Eigennamens bezeichnet und zweitens einen Namensinhalt, welcher Sinn genannt wird. Die Frage nach diesem Sinn des Eigennamens, aber auch Kritik an der vorgeschlagenen Zweikomponenten-Semantik hat zu vielen weiteren Auseinandersetzungen mit Eigennamen geführt (vgl. dazu den Sammelband [WOLF 1993]). [STRAWSON 1958] etwa betont, dass die Referenz einer Gegenstandsbezeichnung nicht kontextunabhängig, also ohne die Verwendung betrachtet werden kann. Zwar bezieht er sich dabei auf Kennzeichnungen, doch gilt dies selbstverständlich auch für Eigennamen. [SEARLE 1958] geht davon aus, dass der Sinn eines Namens in einem Bündel von Kennzeichnungen besteht, die den bezeichneten Gegenstand identifizieren, und wird dafür in [KRIPKE 1972] kritisiert, welches bis heute als Standardwerk gilt. Dieser geht von starren Designatoren aus, die in jeder möglichen Welt, in der sie überhaupt etwas bezeichnen, denselben Gegenstand bezeichnen. Der Bezug von Namen wird dabei durch eine Kausalkette der Kommunikation festgelegt. Als Glieder dieser Kausalkette fungieren dabei grundsätzlich Verwendungen des Namens durch einzelne Sprecher. In Gang gebracht wird diese Kette durch einen Akt der Taufe.

Die eigentliche Namenskunde, die Onomastik, hat traditionellerweise einen Fokus auf die Herkunft und historische Entwicklung von Namen und Namensbedeutungen und ist von daher der Etymologie zuzurechnen. Aber auch die Erforschung des Namensgebrauchs (Namenspragmatik), die Namensstilistik, soziale, politische und rechtliche Fragen zu Namen und Benennungen werden der Onomastik zugerechnet. Darüber hinaus ist eine Vielzahl von Studien zur Morphologie und Syntax von Namen im Allgemeinen, aber auch zu bestimmten Namen, z.B. von Gewässern, Fluren, Haus- und Zuchttieren etc. erarbeitet worden. [EICHLER ET AL. 1995] bieten einen guten Überblick zur Onomastik.

Trotz dieser vielen Studien mangelt es an einer exakten Definition des Phänomens Eigennamen, oder wie es [KALVERKÄMPER 1978: 15] ausdrückt, „eine systematische Namentheorie, die eine aus rein linguistischen Fakten gewonnene Definition des Eigennamens vorlegt und eine Standortbestimmung des Namens im sprachlichen (präziser: im lexikalischen) System bietet, ist m. E. von linguistischer Seite kaum geleistet."

Aus der Perspektive der NER sind wir in der vorteilhaften Lage, die Unschärfe des Begriffs auf die zwei für die Aufgabe relevanten Grundprobleme zu reduzieren und die übrigen Aspekte weiteren Untersuchungen zu überlassen. Zum einen mangelt es an einer klaren Unterscheidung zu den definiten Kennzeichnungen, zum anderen ist es wünschenswert, die Eigennamen von den Gattungsnamen abzugrenzen.

Definite Kennzeichnungen sind Ausdrücken, die wie Eigennamen auf genau ein Objekt referieren, wie z.B. „*der höchste Berg der Welt*". Definite Kennzeichnungen charakterisieren das Referierte im Gegensatz zu Eigennamen. Doch ist dies keine verlässliche Unterscheidung, da manche Namen den Gegenstand nicht nur bezeichnen, sondern auch beschreiben, wie z.B. der Spitzname „*Locke*" für eine Person mit lockigen Haaren, oder „Bellevue" für einen Platz mit einer schönen Aussicht. Eine Diskussion zur systematischen Abgrenzung der beiden Begriffe findet in der Linguistik nicht statt. Die Sprachphilosophie definiert den Eigennamen dadurch, dass er als starrer Designator [KRIPKE 1972] in jeder möglichen Welt denselben Gegenstand bezeichnet. Die NER erfordert die praktische Anwendung dieser Kritierien auf Texte und diese erweist sich als problematisch, wie in Abschnitt 2.2 anhand von Beispielen gezeigt wird.

Zur Abgrenzung gegenüber den Gattungsbezeichnungen sei exemplarisch die Definition des Grammatik-Dudens angeführt. Gemäß [DUDEN 1998: 196] werden mit Eigennamen „ [...] Lebewesen, Dinge u. a. bezeichnet, die so, wie sie sind, nur einmal vorkommen, z. B. bestimmte Menschen, Länder, Städte, Strassen, Berge, Gebirge, Flüsse, Seen, Meere, Fluren und andere Örtlichkeiten, Schiffe, Sterne, menschliche Einrichtungen und geistige Schöpfungen. Mit einem Eigennamen wird also etwas Bestimmtes, Einmaliges benannt; er ist in der Regel einzelnen Lebewesen oder Dingen zugeordnet und gestattet, diese zu identifizieren." Die Eigennamen werden den Gattungsnamen, den Appellativa gegenübergestellt, wobei „unter Gattung eine Gruppe von Lebewesen oder Dingen, die wichtige Merkmale oder Eigenschaften gemeinsam haben" [DUDEN 1998: 196] verstanden wird. Die Definition betont die Referenz von Namen auf einzelne Objekte. Diese so genannte Individualisierungsfunktion von Namen, also die Möglichkeit zur Referenz auf einen

bestimmten Vertreter aus einer Klasse von Objekten, wird gemeinhin als wichtigstes Merkmal von Eigennamen angenommen. Die Eigennamen sind laut [DUDEN 1998: 581] dreigeteilt:

- Die Personennamen, zu denen Vor-, Familien-, Spott- oder Beinamen, die Namen von Göttern, mythische Namen und Völker-, Stammes- und Ortsbewohnernamen gehören.
- Zu den Ortsnamen gehören überregionale und regionale Landschafts- und Raumnamen, Orts- und Siedlungsnamen und Teile davon, wie z.B. Straßen- oder Klosternamen, des Weiteren die Namen von Verkehrseinrichtungen, Gewässern und Flurnamen.
- Die dritte Gruppe von Namen umfasst „unterschiedliche, für die Menschen bedeutsame Erscheinungen der Welt", ist aber „nicht eindeutig einzugrenzen und zu systematisieren" ([DUDEN 1998: 581f]). Unter anderem gehören dazu Tiernamen, die Haustieren von ihren Besitzern gegeben werden, Schiffsnamen, Institutionsnamen, Ereignisnamen (z.B. „*Westfälischer Frieden*") und Produktnamen.

Das Kriterium der Referenz auf bzw. der Zuordnung zu einzelnen Objekten bzw. „einzelnen Lebewesen oder Dingen" ist allerdings nicht verlässlich. Durch die Namen von Völkern wird es auf Gruppen ausgedehnt. Institutionen sind allenfalls auf sehr abstrakte Weise individuelle Objekte. Das Einführen von Produktnamen ist zwar einerseits wünschenswert, da „*VW Golf*", „*Nivea*" intuitiv den Eigennamen zugerechnet werden, widersprechen aber dem Kriterium der unikalen Referenz von Eigennamen, da Produktnamen auf eine ganze Klasse, also alle Autos vom Typ „*VW Golf*", aber nicht auf ein einzelnes Objekt referieren. Allerdings ist die Argumentation irritierend, dass aufgrund der fortwährenden Vervielfachung jedes einzelne Produkt als „ein identisches Element eines Typus" und nicht „als individuelles Element einer Klasse" ([DUDEN 1998: 582]) zu betrachten sei. Zum einen kann sophistisch eingewendet werden, dass auch „weiße Eier" identische Elemente eines Typus seien, zum anderen unterscheiden sich mit demselben Namen bezeichnete Produkte oftmals, wenn sie aus unterschiedlichen Zeiträumen oder Produktionsreihen stammen. Viel eher ist die Zuordnung der Produktnamen zu den Eigennamen dadurch motiviert, dass sie in einem expliziten Benennungsakt entstanden sind. Im Übrigen ist die im Duden vorgeschlagene Klassifikation der Produktnamen keinesfalls unbestritten: Die Grammatik von [EISENBERG 1989] zählt Produktnamen, wenn sie als solche verwendet werden (z.B. „*Er fährt einen Opel*"), zu den Gattungsnamen.

Weitere Namensklassen an der Grenze zwischen Eigen- und Gattungsnamen finden sich in [BAUER 1995]. Ausgehend davon, dass die Onomastik auch beispielsweise Pflanzen- und

Tiernamen untersucht, ohne die damit verbundene Erweiterung des Namensbegriffs zu thematisieren, weist er auf die Unschärfe der referenzsemantischen Definition und auf das Fehlen von grammatischen Kriterien zur Abgrenzung hin. Zum einen müssten anhand der referenzsemantischen Definition auch „*Sonne*“ und „*Mond*“ zu den Eigennamen gezählt werden, zum anderen ist die oft angeführte Artikellosigkeit keinesfalls ein notwendiges Kriterium, weil beispielsweise Eigennamen von Völkern wie „*die Schweizer*“ ebenfalls mit Artikeln benutzbar sind. [BAUER 1995] vergisst zu erwähnen, dass Namen von Institutionen, die Appellativa enthalten, wie z.B. „*Sozialdemokratische Partei Deutschlands*“ einen Artikel erfordern, obwohl sie sowohl das referenzsemantische Kriterium erfüllen. Als Grenzfälle zwischen Eigen- und Gattungsnamen nennt [BAUER 1995] beispielsweise Benennungen aus der Botanik („*Prunus domesticus*“ für Pflaume, die Blume „*Fleißiges Lieschen*“) oder die medizinische Nomenklatur der Krankheitsnamen. Dabei spricht er von „einer Art Eigenname“, von „Appellativa, die den Anschein erwecken, Eigennamen zu sein“, oder von „etwas, was kommunikativ wie ein Eigenname verwendbar und als solches interpretierbar ist“ [BAUER 1995:1619], um das Wesen dieser Übergangsformen zwischen Eigen- und Gattungsnamen zu charakterisieren.

Als weitere Überschreitung der Grenzen zwischen Eigen- und Gattungsnamen werden des Weiteren mehrdeutige Wörter angeführt. Sowohl bei [BAUER 1995], als auch im [DUDEN 1998] werden dazu „*Zeppelin*“ als Luftschiff oder als Nachname und „*Diesel*“ als Motorentyp oder als Nachname angeführt. Allerdings ist dies eher ein definitorisches Problem für Lexikographen, da in einer konkreten Verwendung dieser Wörter immer die eine oder andere Lesart zu erkennen ist, es sei denn, es handelte sich um eine absichtliche Mehrdeutigkeit, wie beispielsweise in Witzen. Das gleiche gilt für die oft angeführte metaphorische Verwendung von Namen, wie z.B. „*Bayreuth ist das* Mekka *der Opernfreunde*“.

Die bildhafte Sprache, also die Verwendung von Metaphern und Metonymien sind hervorstechende Zeugen eines zentralen Mechanismus der menschlichen Sprache, welcher es erlaubt, die Bedeutung eines Wortes ad hoc zu erweitern, nämlich durch die Verwendung in einem bisher ungewohnten Kontext. Dabei ist es auch möglich, dass die ehemals ungewohnte Verwendung konventionalisiert wird und in den regulären Sprachgebrauch aufgenommen wird. Interessanterweise kann bei diesem Prozess problemlos die Grenze zwischen Nomen und Namen überschritten werden, und zwar in beide Richtungen. Vom Namen zum Nomen wird dabei der Wortschatz erweitert, wie in den erwähnten Beispielen „*Zeppelin*“, „*Mekka*“ oder „*Diesel*“. Umgekehrt werden oft Nomen benutzt, um daraus neue Namen zu bilden. Bei

der „Taufe“, also der Benennung neuer Produkte oder Organisationen kann entweder ein neues, bisher unbekanntes Wort geprägt werden (z.B. „*Novartis*“, „*Nivea*“), oder aber, was sehr viel häufiger vorkommt, es kann auf den regulären Wortschatz zurückgegriffen werden (z.B. „*Bund für Umwelt und Naturschutz*“, „*Die Tageszeitung*“). Mit einer sprachhistorischen Perspektive kann dies selbst bei Nachnamen, etwa die von Berufsbezeichnungen abgeleitet sind (z.B. „*Müller*“) oder bei Ortsnamen (z.B. „*Burg-dorf*“, „*Neu-stadt*“) beobachtet werden. Eine detaillierte Auseinandersetzung mit dem Phänomen kann im Rahmen dieser Arbeit nicht geleistet werden. Aber es muss als weitere Quelle für die Unschärfe des Eigennamenbegriffs berücksichtigt werden.

Einen interessanten Ansatz zur Unterscheidung von Gattungs- und Eigennamen benutzen [STEINER 2002: 30]. Dabei werden folgende drei Satzrahmen zum Eigennamentest vorgeschlagen:

Für Eigennamen:	„ARTIKEL GATTUNGSNAME heißt …“
	(1) „*Die Frau heißt Paula*“
	(2) ?„*Das Auto heißt Golf*“
Für Eigennamen:	„Der Name ARTIKEL GATTUNGSNAME ist …“
	(1) “*Der Name der Frau ist Paula*“
	(2) ?„*Der Name des Autos ist Golf*“
Nur für Gattungsnamen:	„ARTIKEL GATTUNGSNAME ist ein/eine …“
	(1) ∗„*Die Frau ist eine Paula“*
	(2) „*Das Auto ist ein Golf*“

Eigennamen müssen in den ersten beiden Testrahmen einsetzbar, in dem letzten jedoch ungrammatisch sein. Im Gegensatz zum Duden klassifiziert das Verfahren Produktnamen wie in (2) nicht als Eigennamen. Die Grammatikalität von (2) ist im ersten und zweiten Testrahmen unklar, wird durch das letzte Kriterium jedoch eindeutig den Gattungsnamen zugewiesen.

Interessant ist der Vorschlag von [HOLZFEIND 1979], der einen engen und einen erweiterten Namensbegriff vorschlägt. Geäußert worden ist die Anregung in der Diskussion um die Einführung der gemäßigten Rechtschreibung im Deutschen, sie harrt jedoch ihrer linguistischen Ausarbeitung. Eine solche Ausarbeitung könnte aus der Sicht der NER folgende Züge tragen: Während der enge Namensbegriff durch die Begrenzung auf Bezeichnungen individueller Dinge und Lebewesen geschärft würde, wäre der erweiterte

Namensbegriff für den Bereich zuständig, der Gattungsnamen und Kennzeichnungen umfasst, denen Namenscharakter bzw. Identifikationsfunktion zugesprochen wird. Ein wichtiges Kriterium solcher Gattungsnamen ist, dass sie durch einen expliziten Einführungs- oder Taufakt in die Sprache gelangen. Dies trifft nicht nur auf Produktnamen, sondern auch auf fachwissenschaftliche Nomenklaturen zu und umfasst damit auch die in Abschnitt 2.2.3 eingeführten Named Entities. Die Kennzeichnungen, die einem erweiterten Namensbegriff zuzurechnen sind, könnten an ihrer Funktion im Text festgemacht werden, wo sie, meist in Abwesenheit des eigentlichen Namens, in namensähnlicher Funktion zur Bezeichnung eines Objekts verwendet werden. Wie im nächsten Abschnitt deutlich wird, mangelt es der NER an klaren Abgrenzungen, so dass diese sicherlich von der Ausarbeitung dieser Skizze in eine anwendbare Definition profitierte.

2.2. Was sind Named Entities?

Der Begriff *Named Entities* stammt nicht aus der Linguistik, sondern aus der automatischen Informationsextraktion. Systeme zur Informationsextraktion werden benötigt, um bestimmte, vorgegebene Informationseinheiten in natürlichsprachlichen Texten zu erkennen. Die Vorgabe der zu erkennenden Informationen erfolgt durch die Beschreibung eines abstrakt formulierten Vorganges in Form eines Templates. Sind beispielsweise Informationen über terroristische Anschläge gesucht, so legt das Template die zu extrahierenden Informationen in der folgenden Form fest:

- *Wer* hat den Anschlag begangen?
- *Wo* fand der Anschlag statt?
- *Wann* fand der Anschlag statt?
- *Wieviele Opfer* forderte der Anschlag?
- *Mit welchen Waffen* wurde der Anschlag durchgeführt?

Im Rahmen der Message Understanding Conferences wurden über mehrere Jahre hinweg Systeme zur Informationsextraktion vergleichend evaluiert. Um die Entwicklung von Systemen zu fördern, die leicht auf neue Sachbereiche bzw. Aufgabenstellungen anpassbar sind, wurden für jede Konferenz neue Vorgänge und damit neue Templates vorgelegt. Außerdem wurden zur detaillierten Analyse der Leistungsfähigkeit der Systeme Teilaufgaben formuliert, die unabhängig von der allgemeinen Extraktionsaufgabe evaluiert wurden. Die

generische Architektur eines Informationsextraktionssystems umfasst folgende Schritte (in Anlehnung an [APPELT & ISRAEL 1999]):

- Tokenisierung: Erkennung von Wortgrenzen, Abkürzungen, Textstrukturen und Formatierungen
- Lexikalische Analyse: POS-Tagging, Kompositabehandlung
- Named Entitiy Recognition: Personen, Firmen, Produkte, Datums- und Zeitangaben, Maßausdrücke
- (Chunk-) Parsing: Analyse von Nominal-, Verbal- und Präpositionalphrasen und koordinierten Phrasen
- Koreferenzauflösung: Koreferenz zwischen Named Entities, Nominalphrasen und Pronomen
- Unifikation (partiell) instantiierter Templates

Bei der MUC-6 ([SUNDHEIM 1995]) wurde erstmalig die Teilaufgabe Named Entity Recognition eingeführt und im [MUC-APPENDIX 1995] detailliert beschrieben:
The Named Entity task consists of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are "unique identifiers" of entities (organizations, persons, locations), times (dates, times), and quantities (monetary values, percentages). ([MUC-APPENDIX 1995: 321])
Die NER-Aufgabe besteht also aus der Erkennung von sprachlichen Ausdrücken, wenn diese entweder Entitäten identifizieren oder Zeit- und Datumsangaben oder Mengenangaben darstellen. Die letzten beiden Aufgaben, die Erkennung von Zeit- und Datumsangaben und Mengenangaben werden im Rahmen dieser Arbeit nicht behandelt. Zwar sind diese für die Informationsextraktion genauso wichtig wie die eigentlichen Entitäten, sie sind aber sehr leicht mit einigen Regeln zu erfassen. Bereits [PALMER & DAY 1997] haben für sechs unterschiedliche Sprachen (chinesisch, englisch, spanisch, portugiesisch, französisch und japanisch) gezeigt, dass die Mengenangaben mit weniger als fünf, die Datums- und Zeitangaben mit weniger als 30 Regeln fast vollständig erfasst werden können. Deswegen betrachten wir diese beiden Aufgaben als gelöst und verstehen unter Named Entities, in Übereinstimmung mit dem größten Teil der einschlägigen Literatur, nur die Erkennung von Entitäten.

Der Begriff der *Entität* ist im Deutschen vorrangig Teil der Fachsprachen der Philosophie, der Informatik und der Politikwissenschaften, während ein Gebrauch in der Allgemeinsprache im Sinne von „Ding“, „Gegenstand“, „Einheit“ kaum zu beobachten ist. Da er im Zusammenhang mit der Named Entity jedoch genau in diesem Sinne verwendet wird, verzichten wir auf den Gebrauch des Wortes Entität oder eine versuchte Übersetzung wie „Benannte Entität“ und verwenden im Folgenden durchgehend den englischen Ausdruck Named Entity (NE).

Was eine NE genau ist bzw. welche sprachlichen Einheiten als NE zu klassifizieren sind, kann nicht übergreifend sondern immer nur in Abhängigkeit von einer bestimmten Anwendung diskutiert werden. Zwar erweckt der Großteil der einschlägigen Literatur den Eindruck, dass die Aufgabe von NER-Systemen immer die Erkennung von Personennamen, Organisationsnamen und Ortsangaben sei. Gewiss kommen Eigennamen dieser Klassen in vielen Textsorten am häufigsten vor, aber es ist durchaus möglich, dass eine Anwendung gar keine Erkennung von Ortsangaben benötigt, dass nicht Organisationen im Allgemeinen sondern nur Fußballmannschaften zu erkennen sind, oder aber dass domänenspezifische Namen, etwa von Himmelskörpern, Produkten, geographischen Phänomenen etc. zu erkennen sind. Zum anderen zeichnet sich der NE-Begriff durch eine große Unschärfe aus, die sich am deutlichsten bei der Korpusanalyse bzw. der NE-Annotation von Texten zeigt. Diese Unschärfe wird in den folgenden Abschnitten im Rahmen der drei wichtigsten Evaluationskampagnen zur NER und den dazu veröffentlichten Korpora bzw. NE-Definitionen beleuchtet, um im Anschluss daran ein Fazit über die Schwierigkeiten des Begriffs der NE zu ziehen.

2.2.1. Die NE-Definitionen in den MUCs

Gemäß der offiziellen NE Definition für die MUC-6 gehören zu den NEs „*proper names, acronyms, and perhaps miscellaneous other unique identifiers*“, welche einer der drei folgenden Kategorien zugehörig sind:

> *„ORGANIZATION: named corporate, governmental, or other organizational entity*
> *PERSON: named person or family*
> *LOCATION: name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc .)”*
> ([MUC-APPENDIX 1995: 322]).

NEs sind also Eigennamen, Akronyme und unter Umständen weitere eindeutige Bezeichnungen, die zu einer dieser Kategorien gehören. Die Kategorie ORGANIZATION

wird zur Verdeutlichung noch um folgende Beispiele erweitert: „*sports teams, stock exchanges, multinational organizations, political parties, orchestras, unions, governmental bodies at any level of importance*" (([MUC-APPENDIX 1995: 326]).

Da die NE-Erkennungsrate auf einem Test-Korpus gemessen wurde, ist die obige Definition um Annotationsrichtlinien erweitert, die festlegen, unter welchen Umständen welche Ausdrücke als eine der drei NE-Kategorien markiert werden. Anhand einiger Beispiele wird im Folgenden die Definition und die Schwierigkeiten der Anwendung der Richtlinien diskutiert Alle Beispiele in den folgenden Abbildung stammen aus den MUC-Richtlinien [MUC-APPENDIX 1995], allerdings wurde die originale Auszeichnung durch eine Platz sparendere ersetzt, in der ORGANIZATION mit „ORG", PERSON mit „PER" und LOCATION mit „LOC" bezeichnet wird.

(1)	*<LOC>North</LOC> and <LOC>South America</LOC>*
(2)	*the <PER>Clinton</PER> government*
(3)	*<ORG>Mips</ORG> Vice President <PER>John Hime</PER >*
(4)	*Ford Taurus*
(5)	*<ORG>Chrysler</ORG> division*
(6)	*the <PER>Kennedy</PER> family*
(7)	*the <LOC>Southwest</LOC>region*
(8)	*<ORG>Temple University</ORG>'s <ORG>Graduate School of Business</ORG>*
(9)	*<LOC>Canada</LOC>'s <ORG>Parliament</ORG>*
(10)	*<ORG>McDonalds of Japan</ORG>*

Abbildung 2.1: Beispiele von NEs aus den MUC-6 Richtlinien

Den Umgang mit Elisionen zeigt (1), was dazu führt, dass „*North*" isoliert als eine NE zu betrachten ist. (2) und (3) verdeutlichen, dass ein Name auch dann als Name zu markieren ist, wenn er den eigentlichen Kopf der NP nur modifiziert. Voraussetzung dafür ist jedoch, dass der Annotator aufgrund des Kontextes oder seines Weltwissens davon überzeugt ist, dass es sich um den Namen einer Organisation oder einer Person handelt. (4) ist nicht zu annotieren, weil es sich dabei um ein Produkt und nicht um die Firma „*Ford*" handelt. (5)-(7) entsprechen zwar der Form von (2), werden jedoch an anderer Stelle eingeführt, um zu illustrieren, dass allgemeine, kleingeschriebene Bezeichnungen, die gemeinsam mit der NE eine NP bilden, nicht zu annotieren sind. (8) zeigt, dass in Possessiv-Konstruktionen das Possessiv-s nicht zur

NE gehört und einzelne NE-Einheiten einer solchen Konstruktion getrennt zu annotieren sind. (9) wird mit derselben Begründung in zwei NEs aufgeteilt. (10) illustriert, dass eine „of LOCATION“ Ergänzung zum Organisationsnamen gehört, wenn dieser nicht vorher explizit endet, etwa durch ein Firmenrechtskürzel (z.B. „*Ltd*“, „*GmbH*“). Das bedeutet, dass (9) in der nur leicht veränderten Form „*Parliament of Canada*“ als eine Einheit zu annotieren ist.
Für die meisten mit NEs bezeichneten Objekte existiert ein vollständiger Name, der beispielsweise bei Organisationen das Rechtskürzel beinhaltet oder bei Ortsangaben die Erweiterungen enthält, die den Ortsnamen eindeutig machen (beispielsweise „Mülheim *an der Ruhr*“). Neben diesem vollständigen Namen existiert eine Vielzahl von Variationen, Kurzformen oder sogar Über- und Spitznamen, mit derselben Referenz. Abbildung 2.2 zeigt Beispiele aus [MUC-APPENDIX 1995], welche die Annotationsrichtlinien für Varianten des vollständigen Namens illustrieren.

(11)	*<ORG>IBM</ORG>*
(12)	*<ORG>Big Blue</ORG>*
(13)	*Mr. <PER>Fix-It</PER>*
(14)	*<LOC>the Big Apple</LOC>*
(15)	*<ORG>Big Board</ORG>*
(16)	*<ORG>IBM</ORG> announced that the company would lay off ...*
(17)	*The Republicans held a rally.*
(18)	*the mayor who built Candlestick Park for the <ORG>Giants</ORG>*

Abbildung 2.2: Varianten mehrteiliger Namen aus den MUC-6 Richtlinien

(11)-(15) zeigen exemplarisch, dass Akronyme und Spitznamen ebenfalls zu annotieren sind. (13) ist der Spitzname für den Vorsteher der CIA, (14) bezeichnet New York und (15) die New Yorker Börse. (16) illustriert, dass Namen von Organisationen zu markieren sind und nicht Nomen wie „*company*“, die auf Organisationen verweisen. Widersprüchlich erscheint, dass in (17) „*Republicans*“ nicht als Variante von „*die Republikanische Partei*“ angesehen wird und deshalb nicht als Organisation markiert wird. Zwar weicht es stärker vom vollständigen Namen ab, als „*Giants*“ in (18), die als Kurzform des Footballteams „*New York Giants*“ als Organisation markiert wird, aber vor dem Hintergrund der in (12)-(15) markierten Spitznamen stellt sich die Frage nach einem verlässlichen Abgrenzungskriterium.

In manchen Fällen weist eine NE gar nicht die Objektreferenz auf, die man a priori annehmen würde. So ist „*Deutschland*“ in „*Deutschland hat die Fußball-WM gewonnen*“, gar keine Ortsangabe, sondern eine metonymische Referenz auf die Organisation „deutsche *Nationalmannschaft*“. Ähnlich ist dies, wenn ein Name in einen weiteren Namen eingebettet ist, wie beispielsweise „*Deutschland*“ in „*IBM Deutschland Entwicklung GmbH*“. Abbildung 2.3 zeigt anhand von Beispielen den Umgang der MUC-6 Richtlinien mit metonymischen und eingebetteten Namen.

(19)	*West Texas Intermediate crude*
(20)	*the Nobel prize*
(21)	*St. Michael*
(22)	*<ORG>Finger Lakes Area Hospital Corp.</ORG>*
(23)	*<ORG>General Hospital</ORG>* - Auszeichnung optional
(24)	*the <ORG>White House</ORG>* - Auszeichnung optional

Abbildung 2.3: Beispiele metonymischer und eingebetteter Namen aus den MUC-6 Richtlinien

In (19) wird „*West Texas*“ nicht als LOCATION markiert, da das „*West Texas Intermediate crude*“ eine Rohölsorte der Firma „*West Texas Intermediate*“ bezeichnet und damit eine zusammengehörende Einheit ist, die nicht weiter unterteilt werden soll. Analog dazu wird auch (20) nicht ausgezeichnet. (21) ist ohne Kontext mehrdeutig und könnte die Nicht-Markierung von Kirchengemeinden oder Kirchen exemplifizieren, dient aber zur Illustration der MUC-Richtlinie, dass Heilige nicht als Personen zu annotieren sind.
(22)-(24) zeigt den Umgang mit Referenzen auf Organisationen, in denen der Bezug durch die Nennung des Sitzes der Organisation vollzogen wird oder die danach benannt sind. Während dies bei (22) aufgrund des Firmenrechtskürzels eindeutig als Organisation zu markieren ist, steht (23) und (24) für eine ganze Reihe mehrdeutiger NEs, unter anderem bei Namen von Hotels, Sport-, Verkehrs- oder Militäranlagen und vielen weiteren Einrichtungen. Solche Nennungen sind in Bezug auf ihre NE-Klasse grundsätzlich mehrdeutig, da mit demselben Ausdruck auf den Ort aber auch auf die Organisation referiert werden kann. (Beispielsweise „*Die <LOC>Messe Essen</LOC> liegt im Norden der Stadt*“ vs. „*Der Sprecher der <ORG>Messe Essen</ORG> …*“). Insbesondere bei Kurzformen (23) oder bei metonymischen Bezügen (24) ist diese Mehrdeutigkeit trotz Berücksichtigung des Kontextes

nicht immer aufzulösen. Deshalb führen die MUC-Richtlinien in solchen Fällen ein optionales Markup ein, was selbstverständlich eine gewisse Unschärfe mit sich bringt.
Für die MUC-7 wurden die NE-Definitionen überarbeitet. Die Modifikationen der MUC-7 Richtlinien ([CHINCHOR & ROBINSON 1998]) gegenüber der MUC-6 Version ([MUC-APPENDIX 1995]) liefert Hinweise darüber, welche Annotationsrichtlinien sich in der Anwendung als schwierig oder als unangemessen erwiesen (vgl. Abb. 2.4). Allerdings zeigt die folgende Diskussion einiger ausgewählter Beispiele, dass damit keinesfalls eine eindeutige Definition erreicht werden konnte.

(25)	*<ORG>Red Sox</ORG>*
(26)	*the <ORG>White House</ORG>*
(27)	*the <ORG>Pentagon</ORG>*
(28)	*Washington* – im Sinne von US-Regierung
(29)	*Uncle Sam*
(30)	*Capitol Hill* – im Sinne von amerikanischer Kongress
(31)	*<LOC>Germany</LOC> invaded <LOC>Poland</LOC> in 1939.*
(32)	*<LOC>Baltimore</LOC> defeated the <ORG>Yankees</ORG> by a score of 4 to 3.*
(33)	*<ORG>Trinity Lutheran Church</ORG>*
(34)	*<ORG>General Hospital</ORG>*
(35)	*The Empire State Building*
(36)	*flew to <LOC>Plymouth Airport</LOC>*
(37)	*<LOC>China</LOC> Film Festival*
(38)	*<ORG>West Texas Intermediate</ORG> crude*

Abbildung 2.4: Beispiele aus den MUC-7 Richtlinien

- Die Markierung der Kurzformen von Namen wurde definiert. Kurzformen sind nur dann zu markieren, wenn es sich dabei um einen Eigennamen handelt und die NE-Verwendung eindeutig zu erkennen ist. Als Beispiel wird (25) als Kurzform der Footballmannschaft „*Bosten Red Sox*“ angegeben. Ein negatives Beispiel zur Abgrenzung fehlt leider.
- Das optionale Markup für metonymische Referenzen auf Organisationen wird obligatorisch, vorausgesetzt, dass die metonymische Verwendung die Aufnahme dieser Lesart in einem Wörterbuch rechtfertigte. Beispiel hierfür sind (26)-(27). Dieses Markup von Nennungen, die auf Metonymie oder anderen Tropen beruhen, wird dadurch

eingeschränkt, dass die benannten Organisationen spezifisch sein müssen, was in (28)-(30) nicht der Fall sei. Allerdings ist die auf einigen Beispielen basierende Abgrenzung keinesfalls eindeutig. So stellt sich die Frage nach dem Status der „*Downing Street 10*“ als Sitz des englischen Premierministers oder nach dem „*Bundeshaus in Bern*“ als Sitz des Schweizer Parlaments.

- Metonyme, die mittels Städte- oder Ländernamen auf politische, militärische, sportliche oder andere Organisationen verweisen, sind als Ortsangaben zu markieren, da eine solche Lesart nicht zur Aufnahme in ein Wörterbuch führt (vgl. (31) und (32)). Allerdings ergibt dies, wie an (32) ersichtlich, erstaunliche semantische Repräsentationen auf der Satzebene.
- Die Ambiguität von Organisationen, die nach dem Gebäude bzw. Hauptsitz benannt sind, wird durch eine durchgehende Markierung als Organisation aufgelöst, auch wenn diese auf die Einrichtung referieren (vgl. (33) und (34)). (35) ist jedoch nicht auszuzeichnen, da es Sitz mehrerer Organisationen ist. (36) wird nicht als Organisation markiert, da Flughäfen explizit den Ortsangaben zugerechnet werden. Dass (35) nicht ebenfalls den Ortsangaben zugerechnet wird, liegt daran, dass nur monumentale Strukturen (exemplifiziert durch „Eiffelturm“) als Ortsangaben zu markieren sind, die in erster Linie als Monument errichtet wurden.
- Die in MUC-6 verfolgte Richtlinie, zusammenhängende Namen, die weitere NEs enthalten, nicht aufzutrennen, wurde aufgegeben. Wenn die längere Sequenz keiner zu markierenden Klasse angehört, wie (37) oder (38), dann können Bestandteile davon markiert werden.

Die Diskussion der Schwierigkeiten der NE-Definitionen in den MUC-Richtlinien könnte anhand vieler weiterer Beispiele vertieft werden. Auch ein Abgleich mit den Definitionen der Automatic Context Extraction Kampagne ACE (z.B. [ACE ANNOTATION GUIDELINES 2004]) als MUC-Nachfolger wäre möglich. Doch soll an dieser Stelle ein Fazit gezogen werden:

- Eine trennscharfe, leicht anzuwendende Definition beliebiger NE-Klassen existiert nicht.
- In den meisten Fällen wird keine explizite Definition angegeben, sondern es werden Richtlinien aufgestellt, die ohne die mitgelieferten Beispiele kaum verständlich wären.

- Die Erstellung der MUC- und ACE-Richtlinien ist äußerst aufwändig und erfordern eine Vielzahl von Überarbeitungen, da jede Annotationsstudie am konkreten Textmaterial zu neuem Klärungsbedarf führt.
- Außerhalb der MUCs und ACEs wurden in keinem NE-Projekt vergleichbare Mühen zur Definition aufgewendet. Ob dies an den erwähnten Schwierigkeiten und/oder dem erforderlichen Aufwand liegt, sei dahingestellt.

2.2.2. NEs im Deutschen - das CoNLL-Korpus

Das einzige frei erhältliche NE-annotierte Korpus deutscher Texte stammt aus dem Shared Task zur NER bei der CoNLL-03 ([TJONG KIM SANG & DE MEULDER 2003]). Die Kampagne diente der Evaluation lernbasierter Systeme zur NER (vgl. dazu Kapitel 4.2). Von den Organisatoren wurde ein Korpus deutscher Zeitungstexte mit NEs annotiert und zum Training und Test der lernbasierten Systeme veröffentlicht. Die NE-Kategorien entsprechen den MUC-Kategorien, also Personen, Ortsangaben und Organisationen und einer zusätzlichen Restgruppe, die alle Eigennamen umfasst, die nicht durch die drei Kategorien abgedeckt sind.
Die manuelle NE-Annotation von deutschen Texten ist sehr gut geeignet, den Begriff der NE genauer zu beleuchten, denn das in den MUC-Richtlinien oft erwähnte Kriterium der Großschreibung von NEs hat für das Deutsche keine durchgehende Gültigkeit. Das Kriterium der Großschreibung führt in den meisten Sprachen mit lateinischem Alphabet implizit dazu, dass alle NEs zu den Eigennamen gehören. Die Unschärfe des Eigennamenbegriffs wird dabei auf die Orthographie ausgelagert, das heißt, alles was außerhalb des Satzanfanges großgeschrieben wird, ist auch ein Eigenname und damit eine potentielle NE. Diese „Vereinfachung“ ist im Deutschen nicht möglich, da nicht nur Eigennamen, sondern alle Nomen großgeschrieben werden.
Über den Annotationsvorgang des CoNLL-Korpus liegen keinerlei Informationen vor. Doch zeigt ein Studium des annotierten Korpus die vielfältigen Probleme, mit denen die Annotatoren konfrontiert waren. Eine systematische Evaluation des Korpus wäre mit einem nicht zu rechtfertigenden Aufwand verbunden, so dass sich die folgenden Ausführungen auf wenige, ausgewählte Beobachtungen beschränken.

Varianten und Kurzformen von Organisationsnamen

Die Varianten und Kurzformen von Organisationsnamen bereiten große Schwierigkeiten, dies kann am Beispiel des Wortes „Sozialamt“ gezeigt werden, welches in spezifischer

Verwendung viermal als Organisation und dreimal nicht markiert ist. Eine Internetrecherche ergibt, dass der offizielle Name der für Soziales zuständigen Stelle der Stadt Frankfurt „*Jugend- und Sozialamt*“ heißt. Diese Nennung kommt in keinem Text vor. Eine konsistente Annotation beruht auf der Festlegung, ob „*Sozialamt*“ als Kurzform des eigentlichen Namens akzeptiert wird. Außerdem muss identifiziert werden, ob sich die Verwendung auf ein bestimmtes, aus dem Kontext zu identifizierendes Sozialamt bezieht oder ob es generisch verwendet wird, wie beispielsweise „*die Deutschen liegen dem Sozialamt auf der Tasche*“. Zweimal taucht auch „*Sozialamt der Stadt Frankfurt*“ auf. Hierbei muss der Annotator entscheiden, ob es sich um den offiziellen Namen der Einrichtung handelt, oder ob hier eine Organisation und eine zusätzliche Ortsangabe zu markieren ist. Dies ist ohne Kenntnis der deutschen Ämterbezeichnungen kaum möglich.

Grundsätzlich stellt sich die Frage, ob eine Nennung von „*Sozialamt*“ überhaupt als NE zu markieren ist. Analog könnte „*Universität*“ als Kurzform der Organisation „*Universität Dortmund*“ annotiert werden, vorausgesetzt, dass die Kurzform ebenfalls auf die „*Universität Dortmund*“ referiert. Allerdings betritt man damit den Bereich der Koreferenzauflösung, welcher nicht zum Aufgabenbereich der NER gehört. Ein Kriterium zur Abgrenzung von Namenskurzform und nominaler Koreferenz ist schwer zu finden:

- Ein mögliches Kriterium wäre, dass eine NE auch ohne Kontext eine eindeutige Referenz erlaubt. Allerdings schließt dies nicht nur „Sozialamt“ aus, da nicht klar ist, um welches Sozialamt es sich handelt, sondern auch „Peter Müller“ und selbst „Frankfurt“ aus, da mehrere Personen mit dem Namen „Peter Müller“ und mehrere Städte mit dem Namen „Frankfurt“ existieren. Auf der anderen Seite wäre „Saarbrücker Universität“ als Organisation zu markieren, obwohl diese offiziell „Universität des Saarlandes“ heißt. Ebenfalls eine eindeutige Referenz liefert die „nördlichste Universität Deutschland“, doch möchte man diese definite Kennzeichnung wohl kaum als Variante der „Universität Flensburg“ als NE markieren.
- Ein weiteres mögliches Kriterium wäre die Forderung, dass eine Variante oder Kurzform Worte oder Wortbestandteile des offiziellen Namens enthält. Allerdings enthalten nicht wenige Organisationsnamen reguläre Nomen, welche die Organisation charakterisieren. Beispiele finden sich in Abbildung 2.5. Es ist fraglich, ob man die Nennung dieser Nomen – selbst wenn sie eindeutig auf eine Organisation verweisen – durchgehend als NE markieren soll. In englischen Texten schreibt man diese Namensbestandteile groß, wenn sie eine Referenz auf den Namen vollziehen und hätte sie dementsprechend als NE zu

markieren. Es besteht aber auch die Möglichkeit, das isolierte Auftreten klein zu schreiben, um die Verwendung als reguläres Nomen deutlich zu machen. Diese Unterscheidung kann im Deutschen auch bei eindeutig auf die Organisation referierenden Ausdrücken sehr schwierig sein. Die Reihenfolge in Abbildung 2.5 ist ein streitbarer Versuch, die „NE-Haftigkeit“ einiger potentieller Namensvarianten absteigend zu sortieren.

	Offizieller Name	**Potentielle Namensvariante** (unterstrichen)
(39)	*Deutsche Bahn AG*	*Die Bahn entlässt Mitarbeiter.*
(40)	*Wuppertal-Institut für Klima, Umwelt, Energie*	*Das Umweltinstitut entlässt Mitarbeiter.*
(41)	*Jugend- und Sozialamt*	*Das Sozialamt entlässt Mitarbeiter.*
(42)	*Universität Dortmund*	*Die Universität entlässt Mitarbeiter.*
(43)	*Deutsche Bank AG*	*Die Bank entlässt Mitarbeiter.*
(44)	*Mayersche Buchhandlung GmbH & Co. KG*	*Die Buchhandlung entlässt Mitarbeiter.*

Abbildung 2.5: Kurzformen und Varianten von Namen

Wie am Beispiel des Wortes „*Sozialamt*“ gezeigt wurde, kann die Folge dieser unklaren Abgrenzung im CoNLL-Korpus in Form von inkonsistenten Annotationen beobachtet werden.

Eine saubere Trennung der hier geschilderten Fälle scheint schwierig. Allerdings darf nicht vergessen werden, dass ein NER-System üblicherweise für eine bestimmte Aufgabe entwickelt wird. Diese Aufgabenstellung spezifiziert die zu erkennenden NEs zusätzlich, so dass viele der hier aufgezählten Schwierigkeiten möglicherweise gar nicht auftauchen. Im CoNLL-Korpus hingegen wurde versucht, eine anwendungsunabhängige NE-Annotation zu erstellen. Des Weiteren wurde versucht, die bereits unscharfen Definitionen für englische Texte ohne viel Aufwand auf deutsche Texte zu übertragen, ohne die Konsequenzen der nicht auf Eigennamen beschränkten Großschreibung zu bedenken.

Bestimmung des Anfangs und Endes einer NE-Sequenz

Durch die nicht auf Eigennamen beschränkte Großschreibung ist im CoNLL-Korpus auch das Problem der genauen Bestimmung des Beginns und des Endes einer NE-Sequenz zu beobachten. Das Korpus enthält 67 Vorkommen des Musters „*Stadt* Ortsangabe“ wie

beispielsweise (45). Dabei ist in 18 Fällen „Stadt“ ebenfalls als Teil der Ortsangabe annotiert. Von den 32 Vorkommen des Musters „*Kreis* Ortsangabe“ hingegen (vgl. (46)) umfasst keine Markierung das vorangehende „*Kreis*“. Dies ist zum einen inkonsistent und zum anderen kaum nachvollziehbar. Schließlich bezieht sich ein einzeln auftretende „Fulda“ automatisch auf die Stadt „Fulda“ und nicht auf den „Kreis Fulda“.

(45)	*Stadt Hofheim*
(46)	*Kreis Fulda*
(47)	*Haltestelle Bleiweißstraße*
(48)	*Parkplatz Zur Untermühle*
(49)	*Kelterei Jung*
(50)	*ökumenischer Weltrat der Kirchen*
(51)	*SPD-Kultusminister*
(52)	*Akkordeon-Musikverein „Heiterkeit“*
(53)	*Volkschor „Frohsinn“ Rödelheim*

Abbildung 2.6: NE-Sequenzen mit unklarem Beginn oder Ende

Im Gegensatz zu den MUC-Annotationen sind auch Adressangaben und Verkehrseinrichtungen als Ortsangaben markiert. Hierzu muss wie in (47) und (48) die Frage geklärt werden, ob die vorangehende Apposition integraler Bestandteil des Namens ist. Die Annotationen im Korpus sind auch hierzu uneinheitlich. Auch bei Organisationen bestehen ähnliche Schwierigkeiten. Im Gegensatz zu den Ortsangaben, bei denen möglicherweise eine definitorische Festlegung Abhilfe verschafft, kann in (49) nur durch die Kenntnis des tatsächlichen Namens entschieden werden, ob die „Kelterei“ Bestandteil des Namens ist. In Sprachen, in denen nur Eigennamen großgeschrieben sind, kommen derartige Schwierigkeiten kaum vor. (50) zeigt, dass noch nicht einmal in journalistischen Texten Verlass auf die Großschreibung von Namen ist. Der offizielle Name der Organisation lautet „Ö*kumenischer Weltrat der Kirchen*“. Die Annotation im Korpus hingegen, hat „*ökumenisch*“ aufgrund der Kleinschreibung nicht als Teil des Namens markiert. (51) zeigt die in deutschen Texten oft zu beobachtenden Bindestrich-Komposita, deren erster Teil ein Organisationsname ist. Diese sind verhältnismäßig konsistent als NE der Restgruppe ausgezeichnet, also als NE, aber nicht der Kategorie Organisation. Allerdings finden sich dennoch Vorkommen wie (51), welches fälschlicherweise als Organisation markiert ist. Ein Ausweg böte die isolierte

Markierung des eingebetteten Organisationsnamens, was jedoch eine Auftrennung des Tokens oder die Teil-Annotation eines Wortes notwendig machte. Ein Argument für die Annotation ist die unstrittige Annotatierung in der äquivalenten Phrase „*Kultusminister der <ORG>SPD</ORG>*“. (52)-(53) zeigen weitere Beispiele für Schwierigkeiten bei Organisationsnamen. Intuitiv ist davon auszugehen, dass ein Name, insbesondere wenn er ungewöhnlich scheint, in Anführungszeichen gesetzt wird. Demnach wäre nur der Inhalt der Anführungszeichen zu markieren. Genauso überzeugend kann jedoch argumentiert werden, dass „*Postsportverein*“ oder „*Volkschor*“ keine gewöhnlichen Appositionen seien, sondern Teile des Namens, welche darüber hinaus auch Eingang in den Eintrag des Vereinsregisters finden. (53) weist zusätzlich die Schwierigkeit auf, dass die Zugehörigkeit der nachgestellten Ortsangabe aufgelöst werden muss.

Eine weitere Auffälligkeit der Annotationen des CoNLL-Korpus ist der inkonsistente Umgang mit neuen, bisher nicht vorkommenden Namensklassen. Dies liegt jedoch nicht an den Besonderheiten der deutschen Sprache, sondern kann bei jeder Annotation beobachtet werden. Das Modell zur Annotierung der NEs, sei es explizit in Form von Richtlinien, oder implizit als Annotations-Verhalten des menschlichen Bearbeiters, wird ständig mit Beispielen konfrontiert, die eine Erweiterung oder Anpassung des Modells erfordern.

(54)	*A 3*
(55)	*B 44*
(56)	*ICE-Trasse Köln-Rhein/Main*
(57)	*Ritterbrunnen*
(58)	*Leunabrücke*
(59)	*Gaststätte „man trifft sich“*
(60)	*das fiktive Christobal*
(61)	*Räuber Hotzenplotz*
(62)	*Shakespeareschen Puck*

Abbildung 2.7: Namen mit unklarem NE-Status

Abbildung 2.7 liefert einige Beispiele von Eigennamen mit unklarem NE-Status:

- Ob Straßennamen (54,55), Eisenbahntrassen (56), Brunnen (57) und Brücken (58) ebenfalls als Ortsangaben zu markieren sind.

- Ob Gaststätten (59) zu den Ortsangaben oder eher – genau wie Firmen – zu den Organisationen zu zählen sind.
- Ob fiktive Orte (60) und fiktive Personen (61, 62) ebenfalls als NE zu markieren sind.
- Wird (61) als Person akzeptiert, ist „Räuber" in diesem Fall Bestandteil des Namens?
- Sollen adjektivische Verwendungen von Namen wie in (62) ebenfalls markiert werden?

Wie die Beispiele deutlich machen, ist die Formulierung von Richtlinien zur Abdeckung der erwähnten Fälle ein aufwändiger Prozess, der darüber hinaus durch jeden neuen Text bzw. die darin vorkommenden Eigennamen potentiell einer Revision bedarf.

2.2.3. Das NE-annotierte Genia-Korpus - NEs in Fachsprachen

Die jüngste Evaluationskampagne zur NER untersuchte die Leistungsfähigkeit von Systemen zur Erkennung so genannter biomedizinischer NEs. Im NE-annotierten Korpus, beschrieben in [KIM ET AL. 2004], sind die NE-Klassen DNA, RNA, Proteine, Zelltyp und Zelllinien („cell-lines") ausgezeichnet. Die NE-Klassen sind aus fachwissenschaftlicher Sicht die Zusammenfassung mehrerer Einträge einer biomedizinischen Ontologie. Insbesondere die Klasse Zelllinien ist eine künstliche Klasse, das heißt, der Ausdruck wird im wissenschaftlichen Diskurs gar nicht verwendet.

	NE	**Klasse**
(63)	*CD-4 coreceptor*	*Protein*
(64)	*T-cell*	*Zelltyp*
(65)	*NF-Kappa B*	*Protein*
(66)	*Unstimulated and 12-O-tetradecanoylphorbol 13-acetate/phytohemagglutinin-stimulated human Jurkat T cells*	*Zelllinie*

Abbildung 2.8: Biomedizinische NEs

Abbildung 2.8 zeigt einige Beispiele biomedizinischer NEs. Aus linguistischer Sicht sind die zu erkennenden NEs als Grenzfälle zwischen Eigen- und Gattungsnamen ([BAUER 1995]) einzuordnen. Viele der NEs enthalten Kombinationen aus Buchstaben und Zahlen, die eine eindeutige Identifizierung des Benannten erlauben und die Klassifikation als Eigennamen rechtfertigt. Allerdings referieren biomedizinischen NEs nicht auf individuelle Instanzen,

sondern auf eine ganze Klasse, d.h. ein Name wie „NF-Kappa B“ bezeichnet alle Proteine von diesem Typ und nicht ein bestimmtes Vorkommen.
Die Annotationen im Korpus sind ohne fundierte Kenntnisse des biomedizinischen Bereiches kaum zu beurteilen. Eine Diskussion der Erkennungsschwierigkeiten dieser oftmals sehr langen und manchmal mehrdeutigen NEs findet sich in Kapitel 3. Der Annotationsprozess ist nicht schriftlich dokumentiert. Laut mündlichen Aussagen bei der JNLPBA-2004 wurde das Korpus von einer Fachkraft mit entsprechenden Kenntnissen annotiert. Inkonsistente Annotationen sind vorhanden, fallen biomedizinischen Laien jedoch nur selten auf. Beispielsweise wurde festgestellt, dass eine Texteinheit doppelt im Korpus vorkommt, die beiden Vorkommen jedoch unterschiedliche Annotationen aufweisen.
Das Korpus steht stellvertretend für eine ganze Reihe fachwissenschaftlicher Anwendungen, die eine NE-Analyse benötigen. Es ist in diesen Fällen noch sehr viel weniger möglich, eine handhabbare Definition der zu erkennenden NEs zu erarbeiten, da Fachkräfte mit der dafür notwendigen Kombination aus Fachwissen und computerlinguistischen Kenntnissen kaum zur Verfügung stehen. Die einzige Möglichkeit bietet hierbei die extensionale Definition der NE-Klassen, die wie im GENIA-Korpus auf einem von einer Fachkraft annotierten Korpus beruht. Dieses Vorgehen kann in einem weiteren Sinne als „beispielsbasierter Wissensaustausch“ zwischen Fachexperten und computerlinguistischen Experten bezeichnet werden. Sie hat den Vorteil, dass der Fachexperte seine Kenntnisse nicht umständlich explizieren muss, sondern diese sehr viel effizienter mittels Beispielen kommunizieren kann. Nachteilig dabei ist, dass auf ein tiefergehendes Verstehen der NE-Klassen von vornherein verzichtet wird und eine darauf basierende Verbesserung der Annotation bzw. der Erkennung der NEs gar nicht erst versucht wird.

2.3. Die Unschärfe des Named Entity Begriffs

In den vorherigen Unterkapiteln wurde der Begriff der NEs aus verschiedenen Perspektiven beleuchtet. In allen Abschnitten hervorstechend die Unschärfe des Begriffs und die vielen Grenzfälle, die mit seiner Anwendung einhergehen. Das ist bei den MUCs zu beobachten, im Verlaufe derer der Begriff „Named Entity“ geprägt wurde und in deren Annotationsrichtlinien immer wieder die Schwierigkeiten der Definition durchscheinen. Weitere Schwierigkeiten werden in den Annotationen des deutschsprachigen CoNLL-Korpus deutlich, in denen der Annotator die im Englischen nicht erforderliche Trennung der Eigennamen von den ebenfalls großgeschriebenen Gattungsnamen zu leisten hat. Weitgehend unergründet blieb das zuletzt

besprochene Phänomen der biomedizinischen NEs. Obwohl die Aufgabe der NER auch in diesem Fall der automatischen Sprachverarbeitung zugeordnet wird, wird es den meisten NE-Experten aufgrund der mangelnden biomedizinischen Fachkenntnisse kaum möglich sein, die NE-Annotationen nachzuvollziehen, hinsichtlich ihrer Konsistenz zu beurteilen oder gar eine Definition zu erarbeiten.
Im Folgenden werden die zentralen Schwierigkeiten des Begriffs der NE zusammengefasst, um anschließend die Konsequenzen für die NER zu skizzieren.

Fazit – die Unschärfe des Named Entity Begriffs

Der Begriff der NE integriert zu weiten Teilen den Eigennamenbegriff. Allerdings ist es der Linguistik bis heute nicht gelungen, eine verlässliche Definition der Eigennamen als eigenständige Wortart zu erarbeiten. Diese Unschärfe überträgt sich direkt auf den Eigennamenbegriff.
Die intuitive Vorstellung über Namen geht davon aus, dass ein Objekt der Welt mit genau einer Wortfolge oder einem Wort bezeichnet wird; welche aus der Sicht der Computerlinguistik eine eindeutige Zeichenkette darstellt. Allerdings werden viele unterschiedliche Zeichenketten dazu benutzt, auf ein bestimmtes Objekt zu referieren. Die menschliche Vorliebe zu Variationen in der Sprache versucht die monotone Wiederholung sprachlicher Einheiten zur mehrfachen Bezugnahme auf das selbe Objekt zu vermeiden und setzt dazu Kurzformen, definite Kennzeichnungen, tropische Bezeichnungen und andere Referenzmechanismen ein. Einige dieser sprachlichen Bezugnahmen haben eine gewisse „Namenshaftigkeit“, sind jedoch weder den Eigennamen zuzuzählen, noch handelt es sich dabei um offizielle Namen wie beispielsweise Taufnamen, rechtsgültige Namen etc. Der NE-Begriff versucht trotz Ermangelung an eindeutigen Kriterien einige dieser sprachlichen Bezugnahmen explizit zu den NEs zu zählen und andere explizit auszuschließen.
Eine weitere Schwierigkeit beruht auf der Klassifikation der Eigennamen in semantische Kategorien. Manche dieser semantischen Kategorien sind Superkategorien, deren Umfang durch eine Menge explizit aufgezählter Subkategorien verdeutlicht wird. Allerdings tauchen in Texten immer wieder Subkategorien auf, deren NE-Zugehörigkeit unklar ist. Beispiele hierfür sind Brunnen, wie etwa der Trevi-Brunnen in Rom, der möglicherweise zu den Ortsangaben zu zählen ist, oder Teile einer Universität, etwa die „Informatik in Duisburg“, die möglicherweise zu den Organisationen zu zählen sind. Darüber hinaus weisen einige semantische Kategorien eine systematische Mehrdeutigkeit auf, die direkt die NE-

Kategorisierung betrifft. Exemplarisch hierfür ist die Mehrdeutigkeit von Organisationsname und Ortsangabe, welche bei Organisationen auftritt, die nach ihrem Sitz benannt sind, wie beispielsweise „*Messe Essen*“ oder die „*Sparkasse Duisburg*“. Je nach Kontext ist die Mehrdeutigkeit klar aufzulösen (vgl. „*ein Brand in der*“ vs. „*die Jahresversammlung der*“) oder bleibt selbst für Menschen ambig. Ein möglicher Ausweg hierzu wird in den ACE-Definitionen ([ACE ANNOTATION GUIDELINES 2004]) aufgezeigt. Die Einführung einer neuen Kategorie „Facilities“, welche Einrichtungen und Anlagen umfasst, deren Name gleichzeitig die dort ansässige Organisation bezeichnet, integriert diese systematische Mehrdeutigkeit. Allerdings ist fraglich, ob dieses Vorgehen die Schwierigkeiten der Zuweisung in wenige Kategorien nicht durch das Problem ablöst dass ständig entschieden werden muss, ob ein bestimmtes Phänomen einer eigenen Kategorie bedarf.

Die unklare Abgrenzung zwischen Eigennamen und Gattungsname verstärkt sich in deutschen Texten, in denen Eigennamen nicht an ihrer Großschreibung zu identifizieren sind. Es ist realistisch, von der Namensnennung, über die Nennung durch Variationen und Kurzformen des Namens, bis zu nominalen Referenzen ein Kontinuum anzunehmen, welches keine klare Abgrenzung zwischen den Bereichen erlaubt. In Sprachen mit exklusiver Großschreibung von Eigennamen erlaubt die Großschreibung als vom Autor eingefügtes externes Kriterium eine Abgrenzung. In deutschen Texten ist der Textautor nicht dafür zuständig diese Unterscheidung zu liefern, vielmehr muss diese Unterscheidung zusätzlich von der NE-Definition bzw. dem NE-Modell geleistet werden.

Eine weitere Schwierigkeit liegt im für die Erkennung von NEs erforderlichen Domänenwissen. Besonderns deutlich wird dies bei Fachtexten. Hierbei sind die Entwickler eines NER-Systems ohne spezielles Domänenwissen nicht in der Lage zu entscheiden und zu verstehen, ob eine sprachliche Einheit als NE zu klassifizieren ist. Allerdings kann dies nicht nur in Fachtexten, sondern selbst in allgemeinen Zeitungstexten beobachtet werden. Auch hier ist es in manchen Fällen schwierig, ohne bestimmte Vorkenntnisse beispielsweise über eine Region, ein Unternehmen oder einen Vereinsnamen zu entscheiden, ob es sich um eine NE handelt.

Konsequenzen aus der Unschärfe des Begriffs

Die aufgezählten Unschärfen der Aufgabe erwecken vielleicht den Eindruck, die Entwicklung von Systemen zur NER sei von vornherein zum Scheitern verurteilt. Wie aber beispielsweise bei den MUC-Evaluationen ([SUNDHEIM 1995], [CHINCHOR & ROBINSON 1998]) deutlich

wird, sind NER-Systeme mit annähernd menschlicher Erkennungsleistung möglich. Es stellt sich deshalb die Frage nach der Relevanz und/oder der Konsequenz der obigen Ausführungen.

- Die NER ist eine anwendungsorientierte Aufgabe, deswegen werden Systeme einzig nach ihrer Erkennungsquote bewertet. Ob ein Modell linguistische Phänomene berücksichtigt oder wie viele der kompliziert erscheinenden Fälle korrekt identifiziert werden, wird meist außer Acht gelassen. Deshalb ist die NER auch weniger in der Computerlinguistik anzusiedeln als vielmehr im Bereich des Language Engineering. Das Vorgehen des Ingenieurs zeichnet sich im Gegensatz zur verstehensorientierten Computerlinguistik dadurch aus, dass die Lösung im Vordergrund steht und phänomenologische Untersuchungen wenig Beachtung erfahren. Der Einzug dieses Ansatzes hat in den 90er Jahren viele Bereiche der automatischen Sprachverarbeitung revolutioniert und hat zu großen Erfolgen bei der Entwicklung leistungsfähiger und anwendungsreifer Systeme geführt. Doch bedarf das Language Engineering der Reflexion und des Bewusstseins über die gleichermaßen faszinierende wie herausfordernde Komplexität der menschlichen Sprache, um diese Erfolge auf weitere Aufgaben der Sprachverarbeitung auszudehnen.
- Um die NER als eigenständige Aufgabe innerhalb der automatischen Sprachverarbeitung zu etablieren, ist ein ganzheitlicher Blick auf mögliche Aufgaben und Herausforderungen notwendig. Eine bestimmte NER-Aufgabe, wie beispielsweise die Erkennung von Ortsangaben, Organisations- und Personennamen in englischen Zeitungstexten, bringt andere Herausforderungen mit sich als beispielsweise die Produktenamenerkennung in finnischen Emails. Ein ganzheitliches Verständnis und Studium des Phänomens dient dazu, die unterschiedlichen Aufgaben auf der Basis der zu erkennenden Einheiten miteinander in Beziehung zu setzen und Lösungsansätze aus der einen Aufgabe auf die Schwierigkeiten einer anderen Aufgabe abzubilden.
- Aus Sicht der anwendungsorientierten Sprachverarbeitung erscheinen die Überlegungen zur Unschärfe des Begriffs der NE sicherlich problemorientiert und zu Recht darf gefragt werden, ob die zum Teil hervorragenden Leistungen in einigen NER-Aufgaben die ganzen Ausführungen Lügen strafen. Aus linguistischer Sicht müssen sich die Ausführungen sogar den Vorwurf der „Polsterstuhl-Linguistik“ gefallen lassen. [FILLMORE 1992:37] hat die sog. “armchair linguists” folgendermaßen karikiert: „He sits in a deep soft armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, 'Wow, what a neat fact!', grabs his pencil, and writes

something down (...) having come still no closer to knowing what language is really like." Auch wenn die Korpuslinguistik bei Fillmore keineswegs besser wegkommt, kann aus anwendungsorienter Sicht argumentiert werden, dass nicht einzelne, schier unlösbare Probleme die Leistung eines Systems bestimmen. Sehr viel zentraler ist die Frage, ob denn ein genügend großer Anteil der in einem Korpus vorkommenden NEs Eigenschaften aufweist, welche die Entwicklung anwendungsreifer NER-Systeme realisierbar erscheinen lässt. Die unscharfen und schwierigen Fälle sind nur dann von Interesse, wenn sie in einem Korpus in großem Umfang vorkommen und die Erkennungsleistung anwendungsrelevanter Einheiten dramatisch nach unten sinken lassen. Ist dies nicht der Fall, so sind sie aus der Sicht der Language Engineers eine Quantité négligeable. Fillmore wäre damit sicherlich nicht einverstanden, streitet er doch gegen die Performanz-orientierte Linguistik und für den Kompetenz-orientierten Ansatz. Dies führte im Falle der NER zwar zu vielen interessanten Fragestellungen, aber kaum zu leistungsfähigen Systemen. Dennoch soll Fillmores Zitat zum Anlass genommen werden, den Polsterstuhl zu verlassen, und den gesamten Fokus auf Lösungen zu konzentrieren, die das Phänomen der Sprache vielleicht nur unwesentlich erhellen, aber Systeme ermöglichen, die Menschen ein nützliches Werkzeug für den Umgang mit Texten sind.

3. Named Entity Recognition

Dieses Kapitel führt in die grundlegenden Aspekte der Named Entity Recognition (NER) ein. In Kapitel 3.1 werden die Aufgabenstellung, die Motivation zur NER und die verwendeten Evaluationsmaße vorgestellt. In Kapitel 3.2 werden ausgehend von den Schwierigkeiten der Aufgabe sprachliche Phänomene diskutiert, welche für die automatische NER eingesetzt werden können. Grundlegend hierfür sind die Begriffe der internen und externen Evidenz. Die Bildung eines Modells zur NER erfordert die Abstraktion sprachlicher Phänomene. Dazu wird die Generalisierung anhand linguistisch motivierter Kriterien der so genannten datenorientierten Generalisierung gegenübergestellt. Dem Einsatz von Listen zur semantischen Kategorisierung, der Problematik der Behandlung von Mehrwortnamen und Aspekten der Texteinheit ist anschließend jeweils ein eigenes Unterkapitel gewidmet.

3.1. Die Aufgabe der NER

Die Aufgabe der NER ist die Markierung derjenigen Wörter und Wortsequenzen einer natürlichsprachlichen Eingabe, die einer der im Vorfeld festgelegten und definierten NE-Kategorien zugehörig sind. Ausgehend von der Diskussion des NE-Begriffs aus Kapitel 2, gehen wir im Folgenden von gegebenen NE-Definitionen bzw. Annotationsrichtlinien aus und lassen vorhandene, definitorische Unschärfen so weit wie möglich außer Acht. Beispiele solcher Kategorien sind LOCATION (Ortsangaben), PERSON (Personennamen) und ORGANISATION (Organisationsnamen). Die Abbildung 3.1 zeigt eine mögliche Ein- und Ausgabe eines NER-Systems.

> ***Eingabe:*** *This was announced in London, Monday, April 16, 2004, by Mr. Peter Hammer, vice president of Decision Support Ltd …*
>
> ***Ausgabe:*** *This was announced in <LOCATION>London</LOCATION>, Monday, April 16, 2004, by Mr. <PERSON>Peter Hammer</PERSON>, vice president of <ORGANIZATION>Decision Support Ltd.</ORGANIZATION> …*

Abbildung 3.1: NE-Annotation mit Markup

Eine alternative, äquivalente Repräsentation zum Markup in Abbildung 3.1 ist die sogenannte Inside-, Outside-, Begin-Notation (IOB–Notation), wie sie in Abbildung 3.2 gezeigt wird.

This/O was/O announced/O in/O London/B-LOC ,/O Monday/O ,/O April/O 16/O ,/O 2004/O ,/O by/O Mr/O. Peter/B-PER Hammer/I-PER ,/O vice/O president/O of/O Decision/B-ORG Support/I-ORG Ltd./I-ORG …

Abbildung 3.2: NE-Annotation mit IOB-Notation

[RAMSHAW & MARCUS 1995] haben diese Notation zur Abbildung von Text-Chunks eingeführt. Chunking beschreibt das Zerlegen von Sätzen in Chunks, wobei Chunk in Abhängigkeit der Anwendung unterschiedlich definiert ist, meist aber NP-, PP- und VP-Chunks umfasst. Im Gegensatz zum Markup wie in Abbildung 3.1 projiziert die IOB-Notation die Annotation nicht auf eine Sequenz, sondern wortweise bzw. auf jedes Token. Hierbei werden alle Tokens mit einer Auszeichnung versehen, die anzeigt, ob ein Token das erste Wort einer NE ist („B-“ plus Klassenname), Teil einer NE ist („I-“ plus Klassenname) oder keiner NE, d.h. der Klasse außerhalb („O“), nachfolgend „keine-NE“ genannt, zugehörig ist. Die IOB-Notation erlaubt die Einzelwort-basierte Annotation und Verarbeitung bei gleichzeitiger Repräsentation von Mehrwortsequenzen.

Die NER hat also die Identifizierung aller NE-Vorkommen einer natürlichsprachlichen Eingabe zum Ziel. Die Auflösung der Referenz oder der Koreferenz, also die Vereinigung aller NEs, die sich auf dasselbe Objekt beziehen, gehört nicht zur Aufgabe der NER. Taucht etwa in einem Text mehrfach der Personenname „Müller“ oder „Peter Müller“ auf, so beschränkt sich die Aufgabe der NER darauf, alle diese Vorkommen zu identifizieren und als NE des Typs Person auszuzeichnen. Um welche Person es sich dabei handelt, beispielsweise spezifiziert durch Geburtsort und -datum oder durch die Sozialversichungsnummer, ist nicht Aufgabe der NER. Auch ist es nicht die Aufgabe der NER zu entscheiden, ob sich diese Nennungen auf eine einzige Person beziehen, oder ob im Text mehrere unterschiedliche Personen mit gleichem Namen erwähnt werden. Diese Aufgaben werden nachfolgenden Verarbeitungskomponenten zugeordnet. Diese möglicherweise künstlich oder sogar unangemessen erscheinende Aufgabenteilung liegt in der Entstehung der NER-Aufgabe im Rahmen der Informationsextraktion begründet (vgl. dazu Kap. 2.2), wobei diese vorrangig zu Evaluationszwecken als Verarbeitungsschritt isoliert wurde. Wie im folgenden Abschnitt zur Motivation gezeigt wird, ist die „klassische“ Aufgabenstellung, die auch im Rahmen dieses Dissertationsprojekts verfolgt wurde, nicht für alle Anwendungen geeignet, welche von den Ergebnissen der NER profitieren können.

3.1.1. Motivation

Die ursprüngliche Motivation zur Definition der NER-Aufgabe bzw. zur Entwicklung von NER-Systemen geht auf die Informationsextraktion zurück (vgl. dazu Kap. 2.2). Systeme zur Informationsextraktion haben die Aufgabe, bestimmte, vorgegebene Informationseinheiten bzw. Beziehungen zwischen diesen in natürlichsprachlichen Eingaben zu erkennen und strukturiert zu erfassen. Hierbei ist die NER elementar, da NEs oftmals die Antworten auf die Kernfragen „Wer“, „Wo“, „Wann“ und „Wieviel“ liefern. In gleicher Weise ist NER Bestandteil von Frage-Antwort-Systemen und von Systemen zur automatischen Textzusammenfassung.

Die Identifizierung der oft für die Einordnung und das Verständnis eines Texts wichtigen NEs unterstützt darüber hinaus weitere Anwendungen. So erlaubt die Anzeige aller in einem Dokument erwähnten NEs einem menschlichen Bearbeiter zwar keine Inhaltsangabe, aber sie ermöglicht oft eine erste Klassifikation des Textes nach Genre, Thema oder eine erste Einschätzung zur Relevanz für eine bestimmte Fragestellung. Allerdings ist für eine derartige Aufgabenstellung nicht mehr die Erkennung jedes Vorkommens einer NE im Text notwendig, sondern es genügt die Identifizierung eines Vorkommens. Darüber hinaus ist in einer solchen Anwendung eine Zusammenfassung bzw. Normalisierung der Varianten und Kurzformen eines Namens erforderlich. Die Information etwa, dass in einem Dokument die NE „*Peter Müller*“ einmal und die NE „*Müller*“ zweimal vorkommen, ist für Menschen redundant. Ein nicht geringer Teil solch redundanter Nennungen kann über Stringgleichheit und einfache Strategien wie „Longest-Match“ reduziert werden, so dass sich die Aufgabenstellung kaum änderte. Allerdings wäre ein anderes Evaluationsmaß sinnvoll als das in dieser Arbeit verwendete (vgl. Abschnitt 3.1.2).

In vergleichbarer Art und Weise kann das Ergebnis der NER auch zum effizienteren Zugriff auf Dokumentensammlungen etwa in Intra- oder Internetsuchmaschinen verwendet werden. So ermöglicht es die Auswahl aller Dokumente, in denen eine oder mehrere bestimmte NEs vorkommen. Interessant ist dies vor allem für mehrdeutige Wortformen wie etwa „Essen“, was sowohl eine Ortsangabe als auch ein gewöhnliches Nomen sein kann. Die NER unterstützt den Informationszugriff hier vor allem durch die semantische Disambiguierung. Auch die Navigation innerhalb größerer Dokumente wird durch die NER unterstützt. So erlaubt ein Dokument, in dem jedes NE-Vorkommen annotiert ist, dem Benutzer gezielt alle Passagen aufzufinden, in denen eine bestimmte NE vorkommt. Diese Form der semantischen Erschließung hat im Übrigen ein klassisches, papiergebundenes Vorbild in Form von

Registern mit Personen- und Ortsnamen, über welche einige fachwissenschaftliche Bücher verfügen. Sowohl die NE-basierte Erschließung von Dokumentensammlungen als auch einzelner, umfangreicher Dokumente profitieren von der Normalisierung der NE-Vorkommen, also der Zusammenfassung von Variationen und Kurzformen eines Namens.
Den Einfluss von NEs auf die Retrievalqualität von Information Retrieval Systemen untersuchten [MANDL & WORMSER-HACKER 2005]. Sie stellten einen deutlichen Zusammenhang zwischen der Anzahl der NEs in der Anfrage und der Qualität des Ergebnisses fest. Allerdings bleibt bisher unklar, ob dieser Effekt durch eine NER-Analyse von Anfragen oder von Dokumenten sinnvoll ausgenutzt werden kann. Darüber hinaus fehlen Untersuchungen, ob der Effekt für alle NE-Klassen in gleicher Weise zu beobachten ist.
Für manche Anwendungen ist nicht die Erkennung und Klassifikation einiger NEs, sondern die Erkennung aller Eigennamen erforderlich. Diese Aufgabe weist viele Parallelen zur NER auf und profitiert von zuverlässigen NER-Verfahren. In Systemen zur maschinellen Übersetzung etwa, verhindert dies die unerwünschte Übersetzung von Namen. Im Rahmen der Syntaxanalyse unterstützt die Erkennung syntaktisch komplexer Namen die angemessene Verarbeitung derselben als eine Einheit und löst damit bereits im Vorfeld Mehrdeutigkeiten etwa beim Attachment von Präpositionalphrasen auf.

3.1.2. Evaluation von NER-Verfahren

Das in dieser Arbeit verwendete und hier beschriebene Evaluationsmaß bezieht sich auf die „klassische“ NER, also auf eine Aufgabe, die als Teilschritt der Informationsextraktion betrachtet wird. Dabei wird davon ausgegangen, dass alle NE-Vorkommen in einem Text zu identifizieren sind und dass die Markierung der gesamten Namenssequenz und nicht das Erkennen der einzelnen Bestandteile maßgeblich ist. In den meisten Evaluationskampagnen werden nur die korrekte Erkennung und Klassifikation des gesamten Namens, nicht aber Teile davon als korrekter Treffer bewertet. Einzig das Evaluationsmaß der MUCs ([SUNDHEIM 1995], [CHINCHOR & ROBINSON 1998]) berücksichtigt auch partielle Treffer, da sowohl die korrekte Klassifikation der einzelnen Wörter als auch die korrekte Erkennung des Beginns und des Endes eines Namens positiv gewertet werden. Das Evaluationsskript des NER Shared Task bei der JNLPBA-2004 ([KIM ET AL. 2004]) liefert neben der Auswertung der exakten Treffer zusätzlich die Angaben über die korrekt erkannten Grenzen der Namenssequenzen, die allerdings nicht in die Gesamtwertung mit einließen.

Die Evaluation von NER-Systemen beruht auf Precision, Recall und dem F-Measure ([VAN RIJSBERGEN 1979]), gemessen auf einem manuell annotierten Testkorpus. Precision, auch Präzision genannt, ist die Anzahl der vom System gefundenen und korrekten NEs geteilt durch die Anzahl aller gefundenen NEs, unabhängig von ihrer Korrektheit.

Precision = (Anzahl NEs korrekt) / (Anzahl NEs gefunden)

Recall, auch Umfang genannt, ist der Anteil der vom System korrekt erkannten NEs an der Anzahl aller NEs im Testkorpus.

Recall = (Anzahl NEs korrekt) / (Anzahl NEs im Korpus)

Das F-Measure oder F-Maß kombiniert Precision und Recall mit folgender Formel:

F-Measure = (2PrecisionRecall)/(Precision+Recall)

Das in [VAN RIJSBERGEN 1979] eingeführte F-Measure erlaubt grundsätzlich auch eine Gewichtung von Precision oder Recall, diese wird jedoch bei der NER Evaluation nicht eingesetzt.
Die beschriebenen Evaluationsmaße können zusammenfassend für alle NEs im Korpus berechnet werden, aber auch getrennt nach NE-Klasse. Das ist insbesondere für selten im Testkorpus vorkommende NE-Klassen informativer.

3.2. Schwierigkeiten und Lösungen für die NER

Sowohl bei der menschlichen als auch bei der automatischen Annotation von NEs besteht die Schwierigkeit in der Klassifikation von unbekannten oder mehrdeutigen Wörtern. Ist das Wort „*Asahikawa*“ unbekannt, so bleibt ohne zusätzliches Wissen unergründet, ob es sich dabei um eine asiatische Sportart, ein Ritual, eine Person oder – wie es tatsächlich der Fall ist – um eine japanische Stadt handelt. Das Wort „Halle“ kann sowohl eine Stadt in Deutschland, ein Nachname oder auch ein gewöhnliches Nomen sein.

3.2.1. Interne und externe Evidenz

Das Wissen über die Zuordnung solcher Wörter in das vorgegebene Kategorienschema der zu erkennenden Namensklassen kann gemäß [McDonald 1996] in *interne* und *externe Evidenz* unterschieden werden.

Interne Evidenz beschreibt das Wissen über das Wort, welches zu klassifizieren ist. Hierzu gehört das Ergebnis einer Lexikonkonsultation, der sog. lexikalischen Ressourcen. Für NER geeignete lexikalische Ressourcen liefern uns Hinweise, die die Einordnung des Wortes in die vorgegebenen Kategorien oder die Kategorie „keine NE“ ermöglichen bzw. unterstützen. Problematisch sind unbekannte Wörter, d.h. Wörter ohne Lexikoneintrag, oder kategorial mehrdeutige Wörter, also Wörter wie beispielsweise „Essen“ als Stadt oder Nomen, die in Bezug auf ihre NE-Klassifikation nicht eindeutig sind. Ebenfalls zur internen Evidenz gehören Informationen, welche die oberflächliche Erscheinung des Wortes liefert, wie beispielsweise die Groß- bzw. Kleinschreibung oder das Vorkommen bekannter Wortbestandteile. So deutet die Endung „-burg“ beispielsweise auf einen Ortsnamen hin.

Externe Evidenz bezieht sich auf Hinweise aus dem Kontext, in dem das zu klassifizierende Wort vorkommt. Der Kontext „die japanische Stadt“ beispielsweise, ermöglicht uns ein nachfolgendes, völlig unbekanntes Wort wie „*Asahikawa*“ als Ortsangabe zu erkennen. Der Kontext ist nicht nur Quelle wertvoller Hinweise für die NER, vielmehr legt der Kontext die Wortbedeutung auch hinsichtlich der NE-Kategorie fest. McDonald führt externe Evidenz im Zusammenhang mit Grammatiken zur NER ein und beschränkt den berücksichtigten Kontext damit implizit auf die Satzebene. Allerdings kann zur NER auch ein größerer Kontext erforderlich sein. Eine Schlagzeile wie „Washington schlägt zurück“ erfordert das Lesen des Artikels oder von Informationen, die das Einordnen des Artikels in den aktuellen Pressediskurs ermöglichen, um zu entscheiden, ob es ich bei dem kategorial mehrdeutigen „Washington“ um einen Boxer oder die Stadt Washington handelt, ob es metonymisch für „amerikanische Regierung“ steht oder ob es die Kurzform einer Firma ist, die beispielsweise „Washington Semiconductor“ heißt. Eine sehr spitzfindige Argumentation könnte sogar behaupten, dass die meisten in dieser Arbeit erwähnten angeblichen NEs aufgrund des Kontextes gar keine NEs seien. Denn die Erwähnung von „*Asahikawa*“ beispielsweise dient gar nicht der Referenz auf die so benannte japanische Stadt, sondern nur als sprachliches Beispiel. Noch deutlicher wird dies durch den Kontext „Köln hat vier Buchstaben“, in dem „*Köln*“ selbstverständlich keine NE ist. Auch wenn diese Argumentation spitzfindig erscheint und für die Entwicklung automatischer NER-Systeme keine praktische Relevanz hat, so

unterstreicht sie noch einmal, dass NER nicht isolierte Wörter klassifiziert, sondern jeweils ein Vorkommen eines Wortes als Bestandteil eines absichtsvoll geäußerten und verständlichen Textbeitrags.

Vergleichbar zum Grundproblem der unbekannten bzw. mehrdeutigen Wörter kann die Voraussagekraft des Kontextes beurteilt werden: Unbekannte Kontexte, also Kontexte über die keine Informationen vorliegen, sind bei der Klassifikation nicht hilfreich. Bekannte Kontexte können die Klassifikation unterstützen, wenn sie auf eine oder mehrere NE-Klassen hinweisen oder diese ausschließen. Viele Kontexte sind jedoch ohne Voraussagekraft, d.h. enthalten keine Hinweise über ein unbekanntes Wort, welches sie umgeben. Eine rein auf externer Evidenz beruhende NER-Methode ist nicht vorstellbar. Selbst im obigen Beispiel „*die japanische Stadt*“ wird auf interne Evidenz zugegriffen, nämlich dass das folgende Wort unbekannt ist, großgeschrieben ist und vermutlich aus einem anderen Sprachraum stammt. Ansonsten würde auch „*Millionen*“ in „*Wir investieren in die japanische Stadt Millionen*“ als Ortsangabe klassifiziert.

3.2.2. Modellorientierte Generalisierung über sprachlichen Einheiten

NER erfordert, wie viele andere Aufgaben der Sprachverarbeitung, einen robusten Umgang mit unbekannten Wörtern oder mit Wörtern, über die nur wenige Informationen vorliegen. Das Phänomen der unbekannten Wörter liegt bei automatischen Systemen an der Endlichkeit der zur Verfügung stehenden Ressourcen. Dieser steht die Komplexität und die Anpassungsfähigkeit der Sprache bzw. die damit verbundene ständige Veränderung der Sprache gegenüber. Die schier unendliche Anzahl von Wörtern, die Erweiterung des Wortschatzes durch Neologismen und der kreative Umgang der Menschen mit Wörtern machen es unmöglich, alle in Texten vorhandene Wörter oder alle Bedeutungsvarianten der bekannten Wörter zu erfassen. Des Weiteren führen auch Fehler in den Primärdaten, beispielsweise Versprecher, Tippfehler oder OCR-Fehler bei Computern, zur Konfrontation mit unbekannten Wörtern. Ein Großteil der unbekannten Wörter sind jedoch Namen und damit korrekte Wörter. Die menschliche Verarbeitung zeichnet sich hier durch große Robustheit aus: Rezipienten gehen sehr souverän mit bisher unbekannten Namen um, die ohne oder mit wenig Erklärung durch den Textproduzenten eingeführt werden.

Informationen über unbekannte Wörter können aus dem Kontext oder über die Generalisierung über Wörtern abgeleitet werden. Um möglicherweise ungenügendes oder gar nicht vorhandenes Wissen über ein Wort anzureichern, wird Wissen aus allgemeineren

Klassen abgeleitet, zu denen das Wort gehört. Ist über das Wort „*lugen*“ beispielsweise nichts Spezifisches bekannt, so kann das Wort durch den Hinweis zur Zugehörigkeit der Klasse der Verben der visuellen Wahrnehmung charakterisiert werden. Die Art und Weise der Klassenbildung ist einzig durch das Vorhandensein mindestens eines distinktiven Merkmals vorgegeben. Ungewohnte Zusammenfassungen wären die Klasse aller aus mehr als sieben Zeichen bestehenden Wörter oder die Klasse aller Wörter, die auf „-m“ enden. Geeigneter erscheint die Zusammenfassung nach für Menschen nachvollziehbaren Kriterien, wie beispielsweise die Klasse der Wörter, die als Berufsbezeichnung fungieren können, oder die Klasse von Komposita, deren Kopf das Wort „*-gesellschaft*“ ist. Merkmale für die Klassenbildung müssen danach beurteilt werden, ob sie ihrem Zweck dienlich sind, d.h. zur effizienteren NER Wissen über Wörter bereitstellen. Aber sie können auch danach beurteilt werden, ob sie plausibel sind, also inwiefern sie mit gängigen Theorien vereinbar sind und das Phänomen zu erhellen vermögen.

Modellbasierte Generalisierung orientiert sich an den gängigen Theorien und Modellen über ein Phänomen und steht im Gegensatz zur datenorientierten Generalisierung, welche sich einzig auf die Auswertung der in den Daten zu beobachtenden Merkmale beschränkt und diese zur Generalisierung nutzt.

Modellbasierte Generalisierung stützt sich bei NER auf sprachwissenschaftliche Theorien. Davon ausgehend bietet sich die Generalisierung über Wörtern nach morphologischen, syntaktischen oder semantischen Kriterien an. Davon ausgehend, dass manche Eigennamen wie etwa „*Möller*“ und „*Müller*“ ähnlich klingen, ist auch eine Generalisierung nach phonologischen Gesichtspunkten denkbar. Allerdings sind keine Ansätze bekannt, die auf der phonetischen und nicht auf der graphemischen Ebene ansetzen. Deshalb wird ein solches Vorgehen hier nicht weiter ausgeführt. Semantisches Wissen bzw. die Zugehörigkeit zu semantischen Kategorien ist von zentraler Bedeutung für NER und wird deshalb in einem eigenen Unterpunkt in 3.2.4 behandelt.

Morphologische Generalisierung

Die morphologische Generalisierung reduziert flektierte Formen eines Wortes auf einen Stamm oder generalisiert über Morpheme. Das Vorkommen von „*-burg*“, „*-dorf*“, -„*stadt*“ beispielsweise deutet auf eine Ortsangabe aus dem deutschsprachigen Raum. Das Verb „*schlafen*“ ist, unabhängig vom Tempus der Verbform, in der dritten Person immer ein starker Hinweis für ein menschliches Subjekt bzw. für einen Personennamen in der Subjektposition.

Aufwand und Nutzen morphologischer Generalisierung hängen von der Sprache, aber auch von den zu erkennenden NE-Klassen ab. Viele NEs, sowohl im Englischen als auch im Deutschen, bleiben abgesehen vom leicht zu erkennenden Genitiv-S unflektiert. Ein Plural ist für die meisten NEs nicht möglich, außer etwa bei Familiennamen (z.B. „*die Kennedys*“), welche laut MUC-Definition ebenfalls zu den NEs gehören ([MUC-APPENDIX 1995]). Allerdings werden Adjektive und manchmal Artikel, die am Kopf einer komplexen NE angebunden sind, flektiert (z.B. „*der Deutschen Bank*“, „*den Grünen*“).

Dennoch variiert die Schreibweise von komplexeren Namen, einerseits durch das Weglassen oder Abkürzen von Bestandteilen, andrerseits durch uneinheitliche Orthographie. Ausgeprägt ist diese Varianz bei biomedizinischen Namen, wo neben einer fehlenden Nomenklatur (vgl. [PEARSON 2001]) auch bei der Schreibweise eine große Vielfalt herrscht. Das Protein „*Interleukin-1 beta*“ taucht z.B. auch als „*Interleukin1-beta*“ oder „*Interleukin 1 Beta*“ auf (vgl. [HANISCH ET AL. 2003]. Das Vorkommen bestimmter Morpheme bringt hier, aber auch für die an Komposita reiche deutsche Sprache, wertvolle Hinweise zutage. Dies gilt neben den oben erwähnten Ortsnamen auch für Organisationsnamen; so deutet das Vorkommen von „*-firma*“ oder „*-verein*“ auf Organisationen hin. Allerdings führt eine reine Reduzierung der Kontextwörter auf die Grundform insbesondere bei Verben zum Verlust spezifischer Informationen. Die Entwicklung einer vollständigen morphologischen Analyse für das Deutsche ist mit großem Aufwand verbunden, so dass ein Rückgriff auf existierende Werkzeuge realistischer scheint. Das [DFKI-REGISTRY] gibt einen Überblick über Werkzeuge zur morphologischen Analyse und deren Erhältlichkeit.

Syntaktische Generalisierung

Eine syntaktische Generalisierung kann auf unterschiedlichen Ebenen erfolgen. Wortklassen (part-of-speech, kurz POS) sind Generalisierungen über dem syntaktischen Verhalten von Wörtern. POS-Tagger weisen auch unbekannten Wörtern, basierend auf Merkmalen der Wortform und des Kontextes, eine Wortart zu. Der POS-Tagger TnT [BRANTS 2000] beispielsweise klassifizierte in englischen Texten 86%, in deutschen Texten 89% der unbekannten Wörter mit der richtigen Wortart. Viele NER-Systeme fürs Englische benutzen die Wortart, um großgeschriebene Wörter am Satzanfang zu klassifizieren: Sind diese weder Nomen, noch Eigenname, noch Adjektive, so handelt es sich mit großer Sicherheit nicht um den Beginn einer NE. Eine Übersicht zu erhältlichen POS-Taggern findet sich im [DFKI-REGISTRY]. Eine vollständige syntaktische Analyse kann ausgehend vom Valenzrahmen des

Verbs durch die Zuweisung der grammatischen Funktion äußerst wertvolle Hinweise zur Verfügung stellen. So ist der Agens in Sätzen mit dem Verb „*produzieren*“ in vielen Fällen ein Firmenname. Erst eine syntaktische Analyse kann dieses Wissen im Deutschen mit seiner verhältnismäßig freien Wortstellung nutzbar machen. Allerdings mangelt es unter anderem dadurch an robusten Parsern, dass diese an der Erkennung von NEs als einer syntaktischen Konstituente scheitern (vgl. [STEINER 2001] oder [RINALDI ET AL. 2004]). Es erscheint deswegen wenig aussichtsreich, NER auf der Basis einer vollständigen syntaktischen Analyse durchzuführen.

Aus syntaktischer Sicht sind NEs meist Nominalphrasen, so dass eine flache syntaktische Analyse, ein sog. NP-Chunking [RAMSHAW & MARCUS 1995] vielversprechend erscheint. Allerdings sind NEs oft Nominalphrasen, deren Kern auf vielfältige Art und Weise erweitert wurde. Problematisch für die NER ist, dass diese Erweiterungen in einigen Fällen zum Namen gehören, in anderen Fällen nicht. Einige Beispiele sollen dies erläutern:

- Vorgestellte Appositionen: Als Teil des Namens in „*Bäckerei Holzmann*“ oder „*Kreis Kleve*“, im Gegensatz zu „Getränkeproduzent *Holzmann*“ oder „Ortschaft *Kleve*“
- Erweiterung durch Adjektive: „*Deutsche Bank*“, „*Evangelische Bildungsstätte Lückendorf*“ als Teil des Namens vs. „deutsche *Commerzbank*“, „evangelische Kirchengemeinde *Schiltach*“. Zwar hilft die kennzeichnende Großschreibung von Adjektiven, die zum Namen gehören, doch selbst in journalistischen Texten wird von der orthographischen Norm in Unkenntnis des Namens abgewichen. Im [FR-CORPUS 1994] findet sich mehrfach der „ökumenische *Weltrat der Kirchen*“, obwohl die Abkürzung „ÖRK“ eindeutig auf die Zugehörigkeit des Adjektivs zum Namen hinweist.
- Nachgestellte Präpositionalphrasen: „*Frankfurt am Main*“, „*Deutsche Gesellschaft für Erziehungswissenschaft*“ als Teil des Namens vs. „*Basel* am Rhein“, „Interessant sind die Kaufangebote der *Jebens Handelgesellschaft* für Dauercamper und Vermieter…“
- Artikel: „*Die Ärzte*“ (eine Rockband) oder „*Die 2 Brüder von Venlo*“ (ein Kaufhaus) mit dem Artikel als Teil des Namens. Das Phänomen ist jedoch selten; oftmals wird gerade die Artikellosigkeit als typisches Kennzeichen von NEs angesehen.

Diese Mehrdeutigkeiten erschweren die Ausnutzung einer flachen syntaktischen Analyse für die NER. Hinzukommt, dass im Deutschen Konstruktionen möglich sind, bei denen die Grenzen zwischen den Konstituenten nur über eine vollständige, syntaktische Analyse oder aber eine vorgeschaltete NER zu erschließen sind („*…dass Microsoft IBM Paroli bieten will*“).

Syntaktische Generalisierung findet deshalb in den meisten NER-Verfahren durch die Berücksichtigung eines minimalen, lokalen Kontextes [THIELEN 1995] statt. Diese Kontexte werden oft Trigger genannt und werden selten nach linguistischen Kriterien, sondern pragmatisch, also anwendungsbezogen eingesetzt.

3.2.3. Datenorientierte Generalisierung über sprachlichen Einheiten

Zwar greifen die behandelten morphologischen und syntaktischen Generalisierungen auf linguistisch anerkannte Sprachmodelle zurück, doch heißt dies nicht, dass diese besonders gut für die NER geeignet sind. Als Alternative bieten sich datenorientierte Verfahren an. Ein für die NER interessanter Teil des in den Wortarten kodierten Wissens ist beispielsweise auch im simplen Merkmal der Großschreibung enthalten. Funktionswörter, die am Satzanfang ebenfalls großgeschrieben sind, sind Wörter der geschlossenen Klasse. Auch ohne Information über die Wortart ist es mit überschaubarem Aufwand möglich, alle oder zumindest die allermeisten Funktionswörter zu kennen. Im Rahmen von ML-Ansätzen zur NER hat sich die Abstraktion über sog. „deterministische Oberflächenmerkmale“ ([BIKEL ET AL. 1997], [WACHOLDER ET AL. 1997], [MIKHEEV 1997]) als nützlich erwiesen. Die Merkmale zeigen beispielsweise an, ob ein Wort nur aus Großbuchstaben besteht, Zahlen enthält, mit einem Punkt endet, und repräsentieren damit auf einer effizient zu verarbeitenden, wortnahen Ebene linguistische Eigenschaften.

Datenorientierte Generalisierung morphologischer Eigenschaften

Auch morphologische Generalisierung kann mittels datenorientierter Verfahren vollzogen werden. Um Hinweise über die Wortart von unbekannten Wörtern zu erhalten, kann Information aus der Endung entnommen werden [BRANTS 2000]. Im Deutschen weisen gewisse Flexionsendungen auf Verbformen hin, (-st,-t, -en etc.), andere Ableitungssuffixe auf Nomen (-ung, -tion, -schaft, -heit etc) oder auf Pluralformenn (-en). Ähnliches gilt für Präfixe wie beispielsweise ver-, zer-, ent-, die auf Verben hindeuten. Wie bereits an den gegebenen Beispielen deutlich wird, sind dies lediglich Hinweise und erlauben selten eindeutige Schlüsse. Allerdings können diese mit weiteren Hinweisen aus dem Wort und dem umgebenden Kontext kombiniert werden und so die Verlässlichkeit der Klassifikation erhöhen.

So genannte Stemming-Verfahren benutzen die Abtrennung erkennbarer Suffixe und Präfixe, um Wortformen ohne Lexikonzugriff auf einen Stamm oder ein Stamm-ähnliches,

informatives Fragment zurückzuführen. Dies kann mit manuell erstellten, einzelsprachenspezifischen Regeln (eingeführt von [PORTER 1980]) oder korpusbasiert (jüngst für NER [KIM 2004]) bewerkstelligt werden. Bei einem korpusbasierten Ansatz wird ein Korpus in einem ersten Schritt nach Stamm-ähnlichen, informativen Fragmenten gefiltert, um diese in einem zweiten Schritt zur Zerlegung der im Korpus vorkommenden Wortformen zu benutzen. Je einfacher die Morphologie einer Sprache, desto verlässlicher funktionieren Stemming-Verfahren. Abhängig von der Reichweite der Affix-Reduktion werden Wortformen auf einen lexembasierten („*liebte, geliebt, liebtest*“ auf „*lieb*“) oder einen auf Lexemfamilien basierten („*Liebe, Liebhaberei, Liebesschmerz, verlieben*“ auf „*lieb*“) Stamm reduziert. Solche Verfahren wurden ursprünglich für das Information Retrieval entwickelt, führen dort zu einem kleineren Suchindex und bei der Reduktion auf Lexeme, nicht aber bei der Reduktion auf Lexemfamilien, zu einer leichten Verbesserung der Retrieval-Qualität (vgl. [BRANTS 2003] für einen Überblick).

Aus Sicht der NER ist eine vollständige Reduktion auf Lexeme oder gar Lexemfamilien nicht wünschenswert: Zum einen tendieren Eigennamen zur Invarianz, so dass die Normalisierung bzw. Berücksichtigung der Endung genügt. Zum anderen verlieren die Wörter des direkten Kontextes dadurch an spezifischer Information. Wird etwa „*singenden*“ und „*singende*“ auf „*singend*“ reduziert, so geht der Hinweis über den Numerus verloren. Dies kann für die NER bedeutsam sein, da Personennamen beispielsweise, abgesehen von ganzen Familien, immer im Singular stehen und deshalb vermutlich öfter auf „*singende*“ als auf „*singenden*“ folgen. Trotzdem ist es aber wünschenswert, dass durch Wortzerlegungen mögliche Zusammenhänge zwischen Wörtern sichtbar werden. Das Wort „*vorsingende*“, sollte, wenn es denn unbekannt ist, mit „*singende*“ in Verbindung gebracht werden. Morpheme wie beispielsweise „*-burg*“ oder „*-heim*“, weisen, wenn auch nicht eindeutig (wie z.B. in „*Altersheim*“) auf einen Ortsnamen hin. Variationen, selbst von bekannten biomedizinischen Namen, sind ohne Wortzerlegung kaum zu erkennen. Dies kann durch die Zerlegung von Wörtern in Buchstaben-N-Gramme erreicht werden. Diese Technik stammt ebenfalls aus dem Information Retrieval und erzielt dort sogar leicht bessere Resultate als Stemming-basierte Verfahren ([BRANTS 2003]). Die Idee ist, Zusammenhänge zwischen Wortformen bzw. Zusammengehörigkeiten aufgrund von übereinstimmenden Ausschnitten aus der Zeichenkette zu entdecken. Dazu werden die Wörter beispielsweise in alle möglichen Buchstabenabfolgen der Länge zwei (Bigramm) oder drei (Trigramm) zerlegt und danach miteinander verglichen. Allerdings kann auch hier zu stark generalisiert werden; so werden beispielsweise, „*gefangen-*

angefangen“, „*Intendant-Intendanten*“, „*Gerichtsbezirks-Bezirksgerichts*“ mit denselben Trigrammen repräsentiert. [MAYFIELD ET AL. 2003] setzten für ein NER-System die Zerlegung in Tri- und Vier-Gramme ein, greifen jedoch gleichzeitig auf die Wortform zu, vermutlich um das Problem der Übergeneralisierung anzugehen. [RÖSSLER 2004a] benutzt eine Zerlegung in Trigramme in Verbindung mit der Wortposition des Trigramms und zusätzlich das letzte Uni- und Bigramm, um die Endung abzudecken. Das Verfahren ist in Abschnitt 5.2.3 beschrieben.

Datenorientierte Generalisierung syntaktischer Eigenschaften

Wie weiter oben bei den Ausführungen über syntaktische Generalisierungen dargelegt, ist es kaum möglich, NER-Systeme mit den Ergebnissen einer vollständigen syntaktischen Analyse eines Satzes zu versorgen. Darüber hinaus kann selbst die flache Analyse eine vorgeschaltete NER erfordern. Die Ausführungen zur datenorientierten Generalisierung syntaktischer Eigenschaften beschränken sich deshalb auf minimale, lokale Kontexte, sog. Trigger, welche weitgehend ohne Rückgriff auf die syntaktische Phrasenebene beschrieben werden.

Eine syntaktisch motivierte Modellierung minimaler, lokaler Kontexte wird kaum versucht. Eher entspricht das Vorgehen dem pragmatischen Sammeln hilfreicher Phänomene und ist insofern datenorientiert. Wortarten-basierte Muster mögen für einige NEs, die etwa durch distinktive Großschreibung als Entities erkennbar sind, Hinweise liefern. Üblicherweise aber müssen lexikalisierte Kontexte verwendet werden, die unter semantischen Aspekten generalisiert werden können. Meist werden sog. Trigger gesammelt, also Wörter, die mit hoher Verlässlichkeit an einer bestimmten Kontextstelle zu einer NE stehen. Beispiele sind Wörter wie „*Dr.*“, „*Herr*“ vor einer Person, Wortfolgen wie „*Sprecher des*“ vor Organisationen oder „*geboren in*“, „*Wahlen in*“ vor Ortsangaben. Verben des Sprechens und der Kommunikation beispielsweise, liefern bei SVO-Sprachen einen guten Hinweis, dass die vor dem finiten Verb stehende Sequenz eine Person oder eine Organisation ist. Personennamen werden oft durch Titel, Funktions- oder Berufsbezeichnungen eingeleitet.

Die datenorientierte Alternative zum manuellen Sammeln und Filtern von Triggern ist die N-Gramm-Modellierung des Kontextes. Die zugrunde liegende Annahme besteht darin, dass es genügt, ein Fenster von N Wörtern zu betrachten, wobei N eine konstante Zahl ist. Gegen N-Gramm Modelle wird oft eingewandt, dass die Fixierung auf einen festen Kontext niemals korrekte Sprachmodelle ermöglicht, da damit mehrere grundlegende Eigenschaften der Sprache missachtet werden: Zum einen gibt es zahlreiche sprachliche Konstruktionen, die die

üblicherweise gesetzten Fenstergrößen von n=3 (ganz selten auch n=4) überschreiten, zum anderen kann über die absolute Position in einem Fenster kaum sprachlich angemessen generalisiert werden. Beispielsweise kann die relative Position des Artikels zum Nomen einer Nominalphrase mit N-Grammen nicht modelliert werden. Diese können direkt aufeinander folgen („*der Vater*“), erst nach einer Sequenz von Adjektiven („*der gute, alte Vater*“) oder weiteren, eingebetteten Konstituenten („*der seinen Sohn liebende Vater*“). In deutschen Sätzen mit Verbendstellung erfassen die N-Gramme der Wörter am Satzanfang das Vollverb nicht. Darüber hinaus sind N-Gramme nicht in der Lage, mit Koordinationen umzugehen. In „*der Polizist und Lebemann Duarte*“ wird der Kontext durch N-Gramme nur teilweise abgebildet, bei „*die Spieler Max, Tilo und Weber*“ fehlt der hilfreiche Kontext „*Spieler*“ gänzlich für die letzte Personennennung.

Trotz dieser offensichtlichen Unvollständigkeit von N-Grammen ist diese Art der Modellierung von großer Bedeutung für jüngere Entwicklungen in der automatischen Sprachverarbeitung. Erst die N-Gramm Modellierung ermöglicht den Einsatz statistischer Verfahren, die beispielsweise bei der Erkennung gesprochener Sprache [RABINER 1989] oder beim automatischen POS-Tagging ([BRILL 1992], [BRANTS 2000]) zu entscheidenden Fortschritten führten. N-Gramme können sowohl auf der Tagebene, d.h. Ergebnisebene, als auch auf der Wort- bzw. Wortklassen-Ebene angewendet werden. Für die NER sind N-Gramme deshalb so bedeutsam, weil sich viele der oben erwähnten Trigger bereits durch ein N-Gramm-Fenster von einem oder zwei Wörtern erfassen lassen.

3.2.4. Semantische Kategorien und lexikalische Ressourcen für die NER

Die Kategorisierung von Wörtern aufgrund ihrer semantischen Eigenschaften kann sowohl für den Kontext als auch für die zu klassifizierenden Wörter Informationen bereitstellen. Was die Wörter aus dem Kontext angeht, so können semantische Kategorien helfen, Phänomene allgemeiner zu modellieren. Anstelle der Beobachtung, dass nach „*Bäcker*“, „*Pilot*“, „*Polizist*“ oder „*Politiker*“ oft ein Personenname folgt, kann angemessener und auch umfassender formuliert werden, dass nach Wörtern der semantischen Kategorie „Berufsbezeichnung“ oft ein Personenname folgt. Gleiches gilt für die zu klassifizierenden Wörter. Die Kombination der semantischen Kategorien „Vorname“ und „Nachname“ erlaubt eine ziemlich verlässliche Erkennung von Personennamen. Obwohl Namen bzw. Namensbestandteile üblicherweise nicht als Bestandteile des Lexikons betrachtet werden und insbesondere das Wissen über spezifische NEs dem Weltwissen oder enzyklopädischen

Wissen zuzurechnen ist, bezeichnen wir im Folgenden alle Kollektionen semantisch kategorisierter Wörter als *lexikalische Ressourcen*.

Geeignete lexikalische Ressourcen müssen zwei Anforderungen erfüllen: Einerseits eine möglichst breite Abdeckung des Wortschatzes, um die Generalisierung über unbekannte Wörter zu unterstützen. Andererseits muss es möglich sein, in den lexikalischen Ressourcen spezifisches Wissen über einzelne Wörter bereitzustellen, um die Mehrdeutigkeit mancher Einträge zu berücksichtigen, wie beispielsweise „*Essen*“ als Ortsangabe oder als gewöhnliches Nomen.

Semantische Kategorien zur Generalisierung

Mit lexikalischen Ressourcen kann das Problem unbekannter Wörter abgemildert werden, wenn diese Auskunft über die semantische Kategorien eines ansonsten unbekannten Wortes erlauben. Existierende, lexikalische Ressourcen wie beispielsweise WordNet [FELLBAUM 1998] bzw. GermaNet [HAMP & FELDWEG 1997] ermöglichen dies für den allgemeinen Wortschatz, enthalten jedoch nur wenige Eigennamen. Solche Ressourcen zielen auf eine Abdeckung des allgemeinen Wortschatzes und müssen deswegen für jede Domäne erweitert werden. Außerdem ist nur ein Bruchteil der in Wordnet enthaltenen Kategorien für die NER von Interesse, so dass üblicherweise spezifisch für die NER entwickelte Ressourcen eingesetzt werden. Lexikalische Ressourcen mit semantisch kategorisierten Eigennamen können semi-automatisch durch den Einsatz von Internetsuchmaschinen und existierenden Listen, wie beispielsweise Branchenverzeichnisse oder geographische Indices, gesammelt werden.

Um auch NEs, die aus mehreren Wörtern bestehen, zu berücksichtigen, bieten sich zwei Strategien an: eine sehr textnahe Repräsentation der gesamten Namenssequenz oder die getrennte Abbildung der Namensbestandteile. Die getrennte Abbildung hat den Vorteil, stärker über einzelne Namen zu generalisieren. Am deutlichsten wird dies wohl bei Personennamen, wo es aufgrund der klaren Syntax „Vorname-Nachname“ wenig Argumente für eine komplette Speicherung der Vornamen-Nachnamen Sequenz gibt. Bei längeren Namen, speziell im Bereich der Organisationen, erschließt sich die Syntax nur mit sehr viel Aufwand und kann darüber hinaus sehr spezifische Erweiterungen für einzelne Organisationsarten erfordern. Aufschlussreich hierzu sind beispielsweise die Studien zu Namen von Bildungseinrichtungen ([LÖTSCHER 1995]) oder von Genossenschaften der DDR ([HELLFRITZSCH 1995]). Die Repräsentation der gesamten Namenssequenz umgeht diese

aufwändige Modellierung zwar, nimmt dafür jedoch aus zwei Gründen eine geringere Abdeckung in Kauf. Der Abgleich eines Textes mit einer Liste von gesammelten Namenssequenzen (vgl. die Abbildung 3.4 in Kapitel 3.2.5) ist entweder sehr aufwendig, wenn auch Variationen der Namen gefunden werden sollen, oder von geringer Ausbeute, wenn nur exakte Treffer berücksichtigt werden. Außerdem vernachlässigt bzw. ignoriert die einfache Abbildung aller Sequenzen allgemeine Regularitäten, welche die Abdeckung nicht erfasster Namen vergleichbarer syntaktischer Struktur ermöglicht. Der Abschnitt 3.2.5 beschäftigt sich eingehender mit der Erkennung und Klassifikation von Mehrwortsequenzen.

Semantische Kategorisierung und Mehrdeutigkeit

Semantische Kategorien sind nicht exklusiv, d.h. ein Wort kann mehreren Kategorien zugehörig sein. Im linguistischen Diskurs wird diese Mehrdeutigkeit sehr viel differenzierter betrachtet. Grundlage ist nicht das hier verwendete, linguistisch nicht definierte „Wort", sondern das Lexem, als Einheit des Lexikons gegenüber der oberflächlichen Erscheinung, der Wortform. Ist ein Lexem mehreren semantischen Kategorien zugehörig, wie beispielsweise das Pferd den Tieren, den Turngeräten oder den Schachfiguren zugerechnet werden kann, so wird von Polysemie gesprochen. Weisen zwei unterschiedliche Lexeme die gleiche Oberfläche, d.h. die gleiche phonetische oder graphematische Erscheinung auf, wie beispielsweise „*Kiefer*" als Baum oder als Körperteil, so spricht man von Homonymie. Obwohl diese Unterscheidung das Phänomen erhellt, mangelt es an exakten Kriterien zu einer verlässlichen Abgrenzung [BUßMANN 1990].

In NER-Systemen – wie auch in vielen anderen Anwendungen der automatischen Sprachverarbeitung – wird aufgrund mangelnder Ressourcen, aber auch weil es für die Aufgabenstellung unerheblich ist, nicht modelliert, ob es sich bei einer Mehrdeutigkeit um Homonymie oder Polysemie handelt. Deswegen wird hier auch weiterhin der linguistisch schwer zu definierende Terminus *Wort* benutzt. Bei Kontextwörtern besteht einzig und allein die Frage nach der Verlässlichkeit des Triggers. Beispielsweise ist es bedeutsam, dass die mehrdeutige Berufsbezeichnung „*Wirt*" weniger verlässlich vor einem Personenamen steht als eine andere Berufsbezeichnung wie „*Polizist*". Ob der „*Wirt*" als Teil einer parasitären Beziehung zwischen Lebewesen jedoch ein Homonym oder ein Polysem zum „*Wirt*" als Betreiber eines Restaurants ist, ist für die Nützlichkeit der externen Evidenz unerheblich.

Ähnliches gilt für die lexikalische Mehrdeutigkeit der NE-Klassifikation, also ob es sich um eine Instanz der vordefinierten NE-Klassen oder um keine NE handelt. Allerdings wirft ein

Blick auf einige Beispiele lexikalischer Mehrdeutigkeiten in der Abbildung 3.3 ein wenig Licht auf die Entstehung von Namen und die auf Mehrdeutigkeiten beruhenden Schwierigkeiten.

	Wort	**Mögliche NE-Kategorie/Lesart**
(1)	*Halle*	Ortsangabe, Person, keine-NE
(2)	*Essen*	Ortsangabe, keine-NE
(3)	*Hinz und Kunz*	Personen, keine-NEs
(4)	*Bauer*	Person, Beruf, Schachfigur
(5)	*Zeppelin*	Person, keine-NE bzw. Artefakt
(6)	*Volkswagen*	Organisation, keine-NE bzw. Produkt
(7)	*MTV*	Organisation, Medienprodukt
(8)	*Philipp Morris*	Person, Organisation
(9)	*Das Weiße Haus*	Organisation, Ortsangabe

Abbildung 3.3: In Bezug auf die NER mehrdeutige Wörter

Die unterschiedlichen Kategorien der Beispiele (1) und (2) sind nicht miteinander in Verbindung zu bringen und sind rein homonymischer Natur. Die Kategorien von (3) stehen metaphorisch in Verbindung, wobei es keinen motivierten Zusammenhang zwischen so heißenden Personen und der Redewendung gibt. Ähnlich liegt der Fall bei (4), obwohl nicht auszuschließen ist, dass Vorfahren eines heute „*Bauer*“ heißenden aufgrund ihrer Tätigkeit „*Bauer*“ genannt wurden. In (5)-(9) sehen wir Kategorien, die metonymisch zueinander in Verbindung stehen. Das Produkt wird aufgrund des Erfinders (5) oder die Firma aufgrund des Produkts (6,7), die Firma nach dem Gründer (8) benannt oder die Organisation nach dem zentralen Standort (9) benannt. Metonymie ist eine zentrale Quelle der Namensgebung und kann dazu dienen, einen neuen Namen mit Informationen anzureichern. In Firmennamen beispielsweise, ist oft ein Verweis auf das Produkt enthalten. Das Entdecken der nicht erkannten Metonymien ist jedoch oft von bescheidenem Erkenntnisgewinn: Dass „*Adobe*“ nach dem „*Adobe Creek*“ benannt ist, welcher nahe dem Haus eines der Gründer vorbeifließt, oder eine Vermutung über die Berufstätigkeit der Vorfahren eines Herrn „*Müller*“, ist kaum rekonstruierbar und von geringem Interesse für die Aufgabenstellung.

Aus Sicht der NER muss die Mehrdeutigkeit von Wörtern und dabei insbesondere die Metonymie als Eigenschaft und produktiver Mechanismus der Sprache angesehen werden.

Ein Modellieren und Zugreifen auf detailliert unterscheidende, lexikalische Mehrdeutigkeiten geht über die Möglichkeiten der NER, aber auch ihre Ansprüche hinaus.

3.2.5. Erkennen und Klassifizieren von Mehrwortsequenzen

Das Erkennen und Klassifizieren von Mehrwortsequenzen bringt einige Besonderheiten mit sich, die bei der Erkennung von Namen, die aus einem Wort bestehen, nicht auftauchen. Diese Besonderheiten sind deshalb bedeutsam, weil die Eigennamenerkennung die Erkennung der gesamten Sequenz erfordert. Je länger eine Sequenz ist, desto größer ist die Chance, einen Bestandteil davon nicht korrekt zu erkennen oder zu klassifizieren. Die Beispiele in 3.4 zeigen Namen, die aus mehreren Wörtern bestehen. Die meisten Mehrwort-NEs bestehen aus einem nominalen Kern, wie beispielsweise „*Fachakademie*“ (4), „*Ministerium*“ (2), „*Association*“ (3) oder aus einem (Orts-)Namen (1) etc., der sowohl davor als auch dahinter erweitert sein kann. Ist ein möglicher Kern identifiziert, so stellt sich die Frage, wie der korrekte Beginn und das Ende der Sequenz gefunden werden kann. Hilfreich sind eindeutige Begrenzungen, wie der Beginn oder das Ende eines Satzes oder Kürzel wie „*e.V.*“ oder „*GmbH*“, die Organisationsnamen üblicherweise beenden. Sog. Stopp-Wörter, d.h. hochfrequente, klein geschriebene Einheiten bzw. Satzzeichen sind unzuverlässige Merkmale. Sie treten zwar niemals isoliert als NE auf, können aber, wie aus den Beispielen ersichtlich, die NE in Verbindung mit zusätzlichen Wörtern erweitern.

(1) *Mülheim an der Ruhr*
(2) *Ministerium für Umwelt, Raumordnung und Landwirtschaft*
(3) *Association pour le bilinguisme en classe dès la maternelle*
(4a) *Fachakademie für Sozialpädagogik der Armen Schulschwestern von Unserer Lieben Frau*
(4b) *Fachakademie für Sozialpädagogik der A. Schulschwestern v.U.L.Fr*
(4c) *Fachakademie für Sozialpädagogik der Armen Schulschwestern v.U.l.F.*
(5) *Thomas-Morus-Schule*
(6) *Theater der Stadt Duisburg*
(7) *Chemieschule Dr. Erwin Elhardt*
(8) *Museum in der alten Schule Pfaffenwiesbach*
(9) *CDU – Christlich Demokratische Union*
(10) *HLfU – Hessische Landesanstalt für Umwelt*

Abbildung 3.4: Mehrwortsequenzen als Namen

Eine weitere Schwierigkeit besteht in ineinander verschachtelten NEs. Zwar sind eingebettete NEs oft durch Bindestriche als nicht eigenständig gekennzeichnet (5), aber diese Regel ist nicht durchgängig (6)-(8). Eingebettete NEs können zu widersprüchlichen Voraussagen des NE-Modells führen. Aber sie führen auch auf der Ebene der NE-Definitionen bzw. der manuellen Annotation zu Unklarheiten. In Abhängigkeit der Anwendung, aber auch von enzyklopädischem Wissen, kann beispielsweise in (8), das Museum oder nur die Schule, in (4a) bis (4c) die Schule oder der Orden als NE betrachtet werden.

Darüber hinaus führt das Weglassen oder Abkürzen wie in (4a) bis (4c) und der Einsatz eines Akronyms in (9) und (10) zu vielfältigen Ausprägungen einer ursprünglichen Form. Uneinheitliche Abkürzungen und Weglassungen machen einen rein listenorientierten Ansatz unmöglich und bedürfen flexiblerer Methoden. Akronyme als Kurzformen langer NEs sind anhand ihrer Oberfläche zwar leicht als Akronyme zu identifizieren, aber die Klassifikation unbekannter Akronyme ist äußerst schwierig. Abgesehen von der auffälligen Aneinanderreihung von Großbuchstaben bieten Akronyme ohne die Auflösung der Abkürzung keine weitere interne Evidenz. Neben lexikalischen Ressourcen mit semantisch kategorisierten Akronymen liefert die in manchen Texten vorkommende Einführung des Akronyms, meist in der standardisierten Form „langer Name (Akronym)“, wertvolle Hinweise zur NE-Kategorie.

3.2.6. NER und Texteinheit

NEs können nicht losgelöst von einem Kontext betrachtet werden. Die bereits eingeführte externe Evidenz als Kontext bezieht sich in ihrer ursprünglichen Form [McDonald 1996] nur auf die Satzebene. Es gibt jedoch auch satzübergreifende Phänomene auf der Ebene des Textes, die eine intensivere Betrachtung verdienen. Für unsere Zwecke ist eine Auseinandersetzung mit den definitorischen Schwierigkeiten des Textbegriffes (vgl. hierzu [Bußmann 1990]) nicht fruchtbar, da die Erkennung oder Definition von Textgrenzen nicht die Aufgabe der NER ist. Die NER ist von der Vorverarbeitung bzw. dem Zustand der zu verarbeitenden Texte abhängig. Davon ausgehend verwenden wir einen pragmatischen Begriff von Text, der auch nicht vom Diskursbegriff abgegrenzt wird. Unter Text wird ein kohärenter Diskurs bzw. eine kohärente Texteinheit verstanden, soweit die Datei- bzw. Dokumentenstruktur eine solche Zuordnung erlaubt. Bei einem Zeitungskorpus beispielsweise ist davon auszugehen, dass die Grenzen zwischen einzelnen Artikeln zu erkennen sind, bei einem automatisch kompilierten Web-Korpus hingegen, kann die

Zuordnung der Datei- und Dokumentstruktur auf kohärente Texteinheiten schwierig sein bzw. eine Vorverarbeitung in Form eines sog. Document Zoning verlangen.

Das für die NER interessanteste Phänomen der Textebene beruht auf der Kohärenz von Texten. NEs sind oft zentrale Elemente eines Textes und werden meist systematisch in den Text eingeführt. Besonders deutlich wird dies bei Nachrichtentexten. Wie an den Beispielen 3.5 deutlich wird, werden selbst bekannte NEs (1)-(3) bei der ersten Nennung durch einen erklärenden lokalen Kontext erläutert. Im weiteren Textverlauf werden die NEs ohne erklärende Kontexte verwendet. Ähnlich verhält es sich mit Lang- und Kurzfassungen von Namen, wie beispielsweise (4) und (5). Bei der ersten Nennung wird die Langform verwendet. Falls ein Akronym vorhanden ist, wird dieses direkt dahinter in Klammern eingeführt. Im weiteren Verlauf wird das Akronym oder eine Kurzform des Namens verwendet. Von dieser Struktur wird meist nur bei NEs in Artikelüberschriften abgewichen.

(1) … *der amerikanische Präsident George W. Bush* …
(2) … *der Automobilhersteller Volkswagen* …
(3) … *US-Hauptstadt Washington* …
(4) *Philipp Holzmann AG* - Kurzform*: Holzmann*
(5) *Fresenius Medical Care* – Akronym*: FMC*
(6) *Deutsche Bank AG* – Kurzform*: Bank*

Abbildung 3.5: NEs in Texten

Diese unterschiedlichen Formen einer NE innerhalb einer Texteinheit stellen zwar eine Schwierigkeit, aber auch eine Ressource für die NER dar. NEs innerhalb von erläuternden Kontexten, aber auch längere NEs, in denen wertvolle Hinweise wie „*AG*“ in (4) vorkommen, können benutzt werden, um Kurzformen mit wenig interner Evidenz und NEs in Kontexten mit geringer Hinweiskraft zu klassifizieren. Textbasiertes Wissen kann auch aufgrund des systematischen Aufbaus mancher Textsorten gewonnen werden. Presseartikel beispielsweise, beginnen meist mit einer Ortsangabe, um dem Leser eine geographische Einordnung des Geschehens zu ermöglichen.

Die Idee, text- bzw. diskursbasiertes Wissen zur Disambiguierung zu benutzten, stammt von [GALE ET AL. 1992]. Ohne Bezugnahme darauf und ohne genauere Hinweise über die Umsetzung findet sich der Einsatz diskursbasierten Wissens auch bei [MCDONALD 1996]. [GALE ET AL. 1992] stießen bei ihrer Arbeit zur automatischen Disambiguierung der Lesart eines mehrdeutigen Wortes auf die sog. „One-sense-per-discourse“-Tendenz. Dieses

Phänomen beschreibt, dass ein mehrdeutiges Wort, wie beispielsweise „*Boxen*“ als Sport oder als Lautsprecher, innerhalb eines Textes eine große Tendenz hat, nur mit einer Lesart vorzukommen. Diese Tendenz gilt auch für NEs und wurde erstmalig von [MIKHEEV ET AL. 1998] ausgenutzt. Allerdings handelt es sich dabei nur um eine Tendenz und wie in (6) gezeigt wird, können gewöhnliche Nomen als Teil einer NE zur fehlerhaften Annahme führen, dass das betreffende Wort innerhalb des Textes immer als NE zu klassifizieren ist. Auch können Personennamen und Ortsnamen als eingebettete Teile von Organisationsnamen zu Fehlern führen. Die eingebetteten Namen verweisen beispielsweise auf den Gründer oder den Hauptsitz einer Firma und gehören also eigentlich zu einer anderen NE-Kategorie. In diesem Fall kann die „One-sense-per-discourse“-Tendenz bei einem erneuten Vorkommen dieser eingebetteten Namen zu fehlerhaften Annahmen führen.

4. Ansätze zur automatischen NER

In diesem Kapitel werden die bisherigen Ansätze zur automatischen NER vorgestellt. Die Anzahl der Publikationen zur NER ist kaum überschaubar. Dies hängt zum einen mit dem großen Bedarf an NER-Systemen als Vorverarbeitung für weitere Anwendungen ab, zum anderen an den populären Evaluationskampagnen. Ganz vorneweg ist hier sicherlich die sechste Message Understanding Conference MUC-6 ([SUNDHEIM 1995]) zu nennen, bei der erstmalig eine eigenständige NER-Evaluation durchgeführt wurde, und die überzeugend zeigen konnte, dass NER mit fast menschlicher Performanz möglich ist. Die nachfolgende MUC-7 ([CHINCHOR & ROBINSON 1998]) wiederum trug maßgeblich dazu bei, dass NER zu einem Feld wurde, auf dem sich die in den 90er Jahren immer wichtiger werdenden ML-Ansätze mit regelbasierten Systemen maßen. Die klare Aufgabenstellung und Evaluation und das verhältnismäßig leichte Erstellen von Test- und Trainingsdaten war sowohl für Forscher aus der statistischen NLP als auch aus dem ML sehr attraktiv und führte zu den Evaluationen über Einzelsprachen-unabhängige NER bei CONLL-2002 ([TJONG KIM SANG 2002b]) und CONLL-2003 ([TJONG KIM SANG & DE MEULDER 2003]), bei denen Lernsysteme auf Korpora unterschiedlicher Sprachen trainiert und miteinander verglichen wurden. Um den immensen Bedarf nach NLP-Unterstützung in der biomedizinischen Forschung mit der Popularität dieser Kampagnen zu kombinieren, motivierte die Lancierung der jüngsten Evaluationskampagne zur NER in biomedizinischen Texten bei der JNLPBA-2004 ([KIM ET AL. 2004]). Nur schon die Vorstellung aller an den erwähnten Evaluationen teilnehmenden Ansätze sprengt den Rahmen dieses Kapitels. Darüber hinaus würde es der Intention nicht gerecht, anhand einer Auswahl der wichtigsten Ansätze und Entwicklungen die Einordnung des in Kapitel 5. vorgestellten eigenen Ansatzes zu ermöglichen.

Die vorgestellten Beiträge werden vorrangig anhand des Modellierungsparadigmas und der berücksichtigten Evidenzen (vgl. dazu Kap. 3.) dargestellt. Abschnitt 4.1 stellt an ausgewählten Systemen das regelbasierten Vorgehen vor. Sehr viel umfangreicher und detaillierter werden in Abschnitt 4.2 die lernbasierten Ansätze besprochen. Von den beschriebenen Lernalgorithmen wird dabei vor allem die Support Vektor Maschine (Abschnitt 4.2.5) hervorgehoben, die auch im eigenen System (Kapitel 5.) eingesetzt wird. Die bisherigen Arbeiten zur Auswertung bzw. Ausnutzung umfangreicher nicht-annotierter Textsammlungen werden in 4.3 diskutiert. Der Einfluss von Eigenschaften der NER-Aufgaben, also der Sprache der Korpustexte und der Besonderheiten der zu erkennenden NE-

Klassen, wird erst in Abschnitt 4.4 systematisch untersucht, und zwar unter der Perspektive der Korpus-adaptiven NER. Damit stehen am Ende dieses Kapitels alle Grundlagen für die Beschreibung des im Rahmen dieser Dissertation entwickelten Korpus-adaptiven NER-Systems (5.) bereit.

4.1. Regelbasierte Ansätze

Als Geburtsstunde der eigenständigen NER kann der regelbasierte Ansatz von [MCDONALD 1996] angenommen werden, der 1993 als Beitrag zum Workshop on Acquisition of Lexical Knowledge from Text präsentiert wurde. Innerhalb des Systems SPARSER erfüllt die Komponente Proper Name Facility (PNF) die NER, wenn auch der Begriff Named Entity erst 1995 in [SUNDHEIM 1995] eingeführt wurde. Selbstverständlich war auch in früheren Systemen die Behandlung von NEs erforderlich, doch ist diese Aufgabe niemals zuvor derartig klar und isoliert (einzig [RAU 1991], [COATES-STEPHENS 1991]) als Aufgabe definiert worden, die einer fokussierten Bearbeitung bedarf. Darüber hinaus kann PNF als *der* Prototyp regelbasierter NER betrachtet werden; die wichtigsten regelbasierten Systeme können alle als Varianten von PNF beschrieben werden.

[MCDONALD 1996] geht von den Schwächen des bisherigen, rein listenbasierten Ansatzes aus und schlägt die Berücksichtigung des Kontextes vor. Erst diese konzeptionelle Unterscheidung der Hinweise auf NEs in die von McDonald benannte *interne* bzw. *externe Evidenz* (siehe auch Kap. 3) ermöglicht es, die bisherigen, listenbasierten Ansätze zu erweitern: Zum einen, um das Problem der prinzipiell niemals vollständigen Listen anzugehen, zum anderen können mehrdeutige Einträge wie „*Philipp Morris*“ als Person oder Organisation nur unter Berücksichtigung des Kontextes aufgelöst werden.

PNF verarbeitet eine Eingabe in drei Schritten: Die Erkennung möglicher Wörter und Wortsequenzen als Namen, die Klassifikation derselben in die Kategorien Person, Firma und Ortsangabe und die Abspeicherung dieses Ergebnisses in einem Diskursmodell. Die Erkennungsstufe beruht auf einem endlichen Automaten, der großgeschriebene Wörter, aber auch Zeichen wie „&“ zu einer Sequenz zusammenfasst. Den Versuch der Aufdeckung der Syntax von Namen mittels kontextfreier Regeln weist McDonald als nicht machbar zurück und bevorzugt die flache Struktur endlicher Automaten. Um die Grenzen der Sequenz genauer zu bestimmen, wird die Ausgabe der endlichen Automaten nachbearbeitet, unter anderem dadurch, dass großgeschriebene Wörter am Anfang eines Satzes entfernt werden, falls sie auf einer Liste mit grammatischen Funktionswörtern vorkommen, aber auch durch das zugrunde

liegende semantische Modell. Dieses erlaubt es beispielsweise, Titel und Funktionsbezeichnung von Personennamen zu trennen. Aufgrund von Namensbestandteilen wie „*Jr.*“, „*Ltd*“, „*Bank*“ oder eingebetteten Ortsangaben wird anschließend versucht, die bereinigte Sequenz einer der Namensklassen eindeutig zuzuweisen. Ist diese auf interner Evidenz beruhende Klassifikation nicht eindeutig, so wird auf externe Evidenz zugegriffen. Diese sind in der Form von einfachen, kontextsensitiven Regeln abgelegt, die beispielsweise formulieren, dass ein Name dann eine Ortsangabe ist, wenn darauf das Wort „*office*“ folgt. PNF ist sogar in der Lage, komplexe Namen mit eingebetteten Klammerstrukturen ("*Richard M. ("tricky Dick") Nixon*“) zu bearbeiten. Anschließend werden die klassifizierten Namen im Diskursmodell abgelegt, unter anderem um nachfolgende Nennungen der Namen zu verarbeiten, die ohne die Klassifikation erleichternde Namensbestandteile wie „*Jr.*“, „*Ltd*“ oder „*Bank*“ daherkommen.

Über die Leistungsfähigkeit von PNF kann nur spekuliert werden, da hierzu einzig die Aussage einer fast 100%-Genauigkeit auf einem nicht genannten Testkorpus erwähnt wird. Auch fehlen Hinweise zum Umgang mit Regelkonflikten, aber auch zum Rückgriff auf das Diskursmodell.

4.1.1. Regelbasierte Systeme bei MUC-6 und MUC-7

Einen hervorragenden Überblick über die Leistungsfähigkeit, Herangehensweise und die Architektur regelbasierter Systeme bietet die NER-Aufgabe im Rahmen der Message Understandig Conferences MUC-6 und MUC-7. Für die MUC-6 ([GRISHMAN & SUNDHEIM 1995]) wurden erstmalig die Teilbereiche der größeren IE-Aufgabe separat evaluiert. Eines der damit verfolgten Ziele war die Entwicklung isolierter Verarbeitungskomponenten, welche aufgrund ihrer Leistungsfähigkeit und Modularität im Rahmen größerer Systeme einsetzbar sind. Die erfolgreichste und prominenteste dieser isolierten Aufgaben ist sicherlich die NER, welche von der Mehrheit der Systeme mit Werten höher als 90% Precision und Recall gelöst wurde.

Generischer Aufbau regelbasierter Systeme

Alle regelbasierten Systeme bei MUC-6 und MUC-7 ähneln in Bezug auf die Architektur der NE-Komponente PNF von [MCDONALD 1996] und arbeiten in den dort beschriebenen drei Phasen:

- Erkennung potenzieller Namenssequenzen anhand von Namenslisten, dem Merkmal der Großschreibung und Negativlisten, d.h. großgeschriebene Wörter, die aber keine NEs sind
- Klassifikation der Sequenzen anhand von Listen aus Namen und Namensbestandteilen und aufgrund hilfreicher, lokaler Kontexte
- Erkennen und Klassifizieren nicht erkannter NEs anhand bereits erkannter NEs

Vor den drei Phasen findet eine Vorverarbeitung statt, in welcher unter anderem der zu verarbeitende Zeichenstrom in Tokens zerlegt wird und der Dokumentenaufbau wie beispielsweise Überschriften, Einleitung etc., berücksichtigt werden kann.
Nach der NER werden die Ergebnisse an die nächsten Aufgaben des MUC-Szenarios weitergereicht. Die stereotype Herangehensweise und der Verzicht auf die Analyse der syntaktischen Satzstruktur liegt vermutlich weniger an dem großen Einfluss des Ansatzes von [McDonald 1996] - der im Übrigen auch selten wegen seiner Architektur zitiert wird - vielmehr scheint das Phänomen den Ansätzen genau diesen Aufbau aufzuzwingen.

Aspekte der Syntax

An der generischen Struktur auffällig ist das Fehlen einer syntaktischen Analyse der Satzstruktur bzw. die Begrenzung der Syntax auf lokale Phänomene. Systeme, deren Aufbau davon abweicht, wie beispielsweise OKI ([Fukumoto et al. 1998]) und LOLITA ([Garigliano et al.1998]), welche die syntaktische Satzstruktur vor der NER parsen, sind Ausnahmeerscheinungen und sind hinsichtlich ihrer Performanz weit abgeschlagen.
Aufschlussreich ist die Verbreitung der flachen Analyse lokaler Phänomene, meist mit endlichen Automaten, anstelle der tieferen und breiteren Analyse mit kontextfreien Grammatiken (CFG). Die Vorteile endlicher Automaten werden besonders bei FASTUS ([Appelt et al. 1995]) und dem von FASTUS inspirierten System der New York University ([Grishman 1995]) hervorgehoben. Der Ansatz von [Grishman 1995] diskutiert darüber hinaus drei Nachteile von CFGs, die nicht nur für die NER, sondern für alle Teilaufgaben der MUC Informationsextraktion gelten:

- CFGs sind viel zu langsam, um ein System in vernünftiger Zeit zu entwickeln und vor allem zu testen
- Globale Parsingergebnisse führen zu Fehlern auf der lokalen Ebene, die einer rein lokalen Analyse nicht unterlaufen wären

- Die Erweiterung bzw. Anpassung einer umfassenden, komplexen Grammatik an eine neue Domäne ist aufwändig, da die Integration domänenspezifischer Konstruktionen üblicherweise zu Konflikten mit der bestehenden Grammatik führt.

Im Vergleich zu CFGs ist die Handhabung flacher, auf lokale Phänomene begrenzter Regeln weniger komplex, erfordert aber dennoch eine Strategie, die dem Umstand Rechnung trägt, dass unterschiedliche Regeln, aber auch unterschiedliche Reihenfolgen der Regelanwendungen, zu unterschiedlichen Ergebnissen führen können.

Eine strikte Reihenfolge in Form einer festen Anzahl von Verarbeitungsphasen verfolgen die kaskadierenden Ansätze von [APPELT ET AL. 1995] und [GRISHMAN 1995], bei welchen endliche Automaten jeweils die Ausgabe der vorherigen Stufe verarbeiten und an die nächste Stufe weiterreichen. LaSIE-II ([HUMPHREYS ET AL. 1998]) arbeitet mit kontextfreien Regeln lokaler Phänomene und verfolgt bei Mehrdeutigkeiten eine konservative Parsingstrategie, welche immer die kürzeste semantisch interpretierbare Konstituente (NP, VP, S oder RelativSatz) bevorzugt. Dies führt zu einer Art Chunk-Parsing, die laut [HUMPHREYS ET AL. 1998] vermutlich problemlos in eine Kaskade endlicher Automaten umgewandelt werden könnte.

In NetOWL™ ([KRUPKA & HAUSMANN 1998]) und FACILE (Fast and Accurate Categorisation of Information by Language Engineering, [BLACK ET AL. 1998]) werden den Regeln Gewichte zugewiesen, wobei sich die Analyse mit dem höchsten Gewicht durchsetzt. Leider fehlt bei beiden Systembeschreibungen Genaueres sowohl über die Herkunft dieser Gewichte als auch zum Einsatz derselben. Eine Darstellung in ([KRUPKA & HAUSMANN 1998]) impliziert, dass es sich auf der Ebene der lexikalischen Einheiten um polare Bits (+/-) handelt, wobei ein negatives Bit zur Zurückweisung einer NE-Klassifikation führen soll, wenn ein Wort mit negativem Bit Teil des Namens ist. In FACILE haben Regeln einen sog. Certainty-Wert, dessen Default-Wert 0 ist und dessen Wertebereich auf Werte zwischen 1 und -1 verändert werden kann. Bei konkurrierenden Regeln werden höherwertige Regeln und Regelkombinationen bevorzugt.

Verwendete lexikalische Ressourcen

Neben den Regeln zur Modellierung von Namen und Kontexten ist eine Diskussion der lexikalischen Ressourcen von großem Interesse. Leider ist eine vergleichende Diskussion über die teilnehmenden Systeme nur in Ansätzen möglich. Dies liegt zum einen an den

spärlichen Beschreibungen der verwendeten lexikalischen Ressourcen, zum anderen an der Schwierigkeit einer qualitativen Beurteilung lexikalischer Ressourcen.

Wichtige Eigenschaften lexikalischer Ressourcen sind deren Organisation in semantische Kategorien und deren Umfang bzw. Abdeckung. Was die Organisation in semantische Kategorien betrifft, so wären Fragen der Granularität und der Beschaffenheit der Kategorien von Bedeutung. Leider gibt keines der teilnehmenden Systeme vertiefend Auskunft darüber, was darauf schließen lässt, dass diese Ressourcen sehr pragmatisch, ohne leitendes Theoriemodell zusammengestellt werden. Vereinzelt lässt sich eine eher grobkörnige Organisation erahnen, beispielsweise bei der ersten Version von LaSIE ([WAKAO ET AL. 1996]) wird „*stocks*“ als „organisation-related thing“ bezeichnet, welches dazu benutzt wird, in „*Ericson stocks*“ den ersten Teil als Organisation zu erklären. Darüber hinaus wird manchmal auf den Einsatz von negativen Kategorien hingewiesen (z.B. [KRUPKA & HAUSMANN 1998]), also Wörter, die ansonsten irrtümlich als NE einer bestimmen Kategorie erkannt werden. Was die Disambiguierung der Wörter am Satzanfang betrifft, so wird erstaunlich selten auf existierende POS-Tagger zurückgegriffen. [GRISHMAN 1995] verwendet ein kommerzielles Tool von BBN, [BLACK ET AL. 1998] ein kommerzielles Tool von InXight und [HUMPHREYS ET AL. 1998] den Brill-Tagger ([BRILL 1992]).

Was den Umfang der verwendeten lexikalischen Ressourcen betrifft, so ist eine quantitative Analyse von geringem Interesse, da bei Listen von mehreren Tausend Einträgen ein Großteil der in den Listen geführten Einträge weder in Trainings- noch in Testdaten jemals vorkommt und deshalb faktisch ohne Einfluss bleibt. In den Listen vorkommende, in Bezug auf die NE-Klasse mehrdeutige Einträge können den Wert einer Liste erheblich mindern. In FACILE ([BLACK ET AL. 1998]) wurde die von den MUC-Veranstaltern veröffentlichte Liste von Ortsnamen durch eine eigene manuell erstellte Liste ersetzt, da die MUC-Liste zuviel irrelevante und vor allem zuviel mehrdeutige Einträge enthielt, die das System irreführten. Sehr aufschlussreich sind die in [KRUPKA & HAUSMANN 1998] veröffentlichen Experimente zum Umfang lexikalischer Ressourcen. Zur Entwicklungszeit testeten sie den Einfluss der Größe der lexikalischen Ressourcen und stellten fast keinen Einfluss bei der Reduktion der Liste von 110.000 (F-Measure: 91.6) auf 25.000 (F-Measure: 91.45) und einen geringen bei der Reduktion auf 9.000 Einträge (F-Measure 89.13) fest. Gleichzeitig führte die Erweiterung um 42 (!) domänenspezifische Einträge bei der Kategorie Organisation sowohl bei Precision als auch bei Recall zu einer Verbesserung von 6 Punkten, bei den anderen Kategorien zu einer Verbesserung von bis zu 3 Punkten.

4.1.2. Regelbasierte NER für deutsche Texte

Es existieren nur zwei Ansätze zur regelbasierten NER für deutsche Texte, welche sich weitestgehend an dem Aufbau der in 4.1.1 vorgestellten Systeme orientieren, jedoch aufgrund der spezifischen Schwierigkeiten in deutschen Texten deutlich schlechtere Ergebnisse erzielen.

In [PISKORKSI & NEUMANN 2000] bzw. [NEUMANN & PISKORKSI 2002] wird das IE-System SPPC vorgestellt, welches für deutsche Texte neben Chunk-Parsing und der Annotation grammatischer Relationen auch NER durchführt. NER findet nach dem POS-Tagging statt, welches auf einem eigenen Lexikon beruht und durch Regeln optimiert wird. Die Regeln werden von einen Brill-Tagger ([BRILL 1992]) vorgeschlagen und manuell kontrolliert. Der Gedanke, die NER in das Chunk-Parsing zu integrieren, wurde deshalb verworfen, weil die in NEs oftmals vorkommenden Punkte die für das Chunk-Parsing erforderliche Satzgrenzenerkennung erschweren.

Die NEs werden von gewichteten, endlichen Automaten unter Berücksichtigung des Kontextes extrahiert, in einem „dynamischen Lexikon“ abgelegt und zur Erkennung nachfolgender, nicht entdeckter NEs benutzt. Zwar wird der Einsatz der Gewichtung der endlichen Automaten detailliert beschrieben, doch finden sich keine Angaben zur Gewinnung der jeweiligen Gewichte. Sehr interessant wären Informationen über Art und Umfang der lexikalischen Ressourcen. Erwähnt sind kleine, NE-spezifische Ressourcen, unter anderem die Liste der 50 größten Firmen. Dennoch erreicht der Ansatz, evaluiert auf 20.000 Wörter ([NEUMANN & PISKORKSI 2002]), sehr gute Werte (F-Measure Person 88; F-Measure Organisation 79; F-Measure Ortsangaben 81).

In [VOLK & CLEMATIDE 2001] wird eine isolierte NER-Komponente präsentiert, welche auf deutschen Texten der Computerbranche arbeitet. In ihrem System ist die NER deshalb dem POS-Tagging vorgeschaltet, da sich beim Vergleich verschiedener Tagger für das Deutsche [VOLK & SCHNEIDER 1998] die Unterscheidung von Nomen und Namen als große Fehlerquelle herausgestellt hat. Das System benutzt umfangreiche lexikalische Ressourcen, beispielsweise 16000 Vornamen, mehrere tausend Ortsnamen und greift darüber hinaus auch auf Gertwol [HAAPALAINEN & MAJORIN 1994] zu, ein kommerzielles morphologisches Analysesystem, das ebenfalls kategorisierte Eigennamen enthält. Mithilfe dieser Ressourcen und einer überschaubaren Anzahl Regeln werden NEs erkannt bzw. „gelernt“, wie es die Autoren nennen. Diese Einträge werden gefiltert und wie üblich auf nicht erkannte Namen angewendet. Einzigartig an dem Ansatz ist allerdings, dass die gelernten Einträge nicht für

eine Texteinheit, d. h. einen Artikel, gültig sind, sondern immer nur für die 15 folgenden Sätze. Wird ein Name jedoch innerhalb dieser 15 Sätze noch mal erwähnt, so wird die Gültigkeit um weitere fünf Sätze erhöht. Diese Technik bietet sich vor allem für Texte an, deren Dokumentstruktur nicht zugänglich ist. Zwar hat das System Zugriff auf die Artikelgrenzen, aber das dynamische Vergessen von Einträgen wird dennoch angewendet, da die „One-sense-per-discourse"-Tendenz ([GALE ET AL. 1992]) laut [VOLK & CLEMATIDE 2001] häufig nicht zutrifft. Der Ansatz erreicht, evaluiert auf 990 Sätzen, mit SPPC vergleichbare Werte (F-Measure Person 89; F-Measure Organisation 78; F-Measure Ortsangaben 85).

4.2. Ansätze des Maschinellen Lernens

Im Gegensatz zu den regelbasierten Systemen, welche auf rein intellektuell erzeugtem Wissen beruhen, integrieren die ML-Ansätze eine überwachte Lernerkomponente. Diese dient dazu, anhand vorgegebener Beispiele ein Modell zur NER abzuleiten oder ein solches zu optimieren. Die Idee dahinter ist, die arbeitsaufwändige, manuelle Kombination der unterschiedlichen Evidenzen beim Schreiben von Regeln durch einen Prozess zu ersetzen, der die Evidenzen anhand vorgegebener Beispiele automatisch in einem Modell kombiniert.

Eine wichtige Voraussetzung zum Einsatz von ML-Verfahren ist die Einzelwort- bzw. Token-basierte Perspektive auf die NER: Die Erkennung eines möglicherweise aus mehreren Tokens bestehenden Namens wird in mehrere tokenbasierte Einzelaufgaben zerlegt, bei denen jedem Token im Text eine Markierung zuzuweisen ist, welche angibt, ob das Token eine NE, Bestandteil einer NE-Sequenz oder ein gewöhnliches Wort ist. Erst diese Vereinfachung ermöglicht den Einsatz derjenigen Lernverfahren, die anhand des POS-Taggings bewiesen, dass ML-Ansätze durchaus ernstzunehmende Alternativen zu den herkömmlichen, regelbasierten Systemen darstellen.

Zur Repräsentation hat sich mittlerweile die bereits in Kapitel 3.1 eingeführte IOB-Notation durchgesetzt. Die Inside/Outside/Beginn-Notation kann in verschiedenen Varianten eingesetzt werden. Das Beginn-Tag wird eigentlich nur benötigt, um zwei direkt aufeinander folgende NE-Sequenzen voneinander zu trennen. Aber es können auch alle Wörter am Beginn einer NE-Sequenz mit dem „B-NE-Klasse"-Tag ausgezeichnet werden. Allerdings ist das Phänomen der direkt aufeinander folgenden NEs derselben Kategorie äußerst selten, so dass in den meisten Fällen gar keine B-Klasse gelernt wird bzw. der Einfluss davon kaum zu erkennen ist. Auch kann anstelle eines Beginn-Tags ein End-Tag verwendet werden, so dass

die Sequenzen gewissermaßen entgegen der Leserichtung ausgezeichnet werden. Allerdings sind keine solchen Ansätze bekannt. [TJONG KIM SANG & VEENSTRA 1999] haben den Einfluss der unterschiedlichen Repräsentationen auf das Chunk-Parsing untersucht und nur einen minimalen Einfluss auf die Ergebnisse ihres Verfahrens festgestellt.

Die zunehmende Bedeutung und Evolution des ML für NER wird anhand der durchgeführten Evaluationskampagnen deutlich. Die ersten ML-basierten Systeme sind zu Zeiten der [MUC-6 1995] und [MUC-7 1998] entwickelt worden und erzielten schwächere Ergebnisse als regelbasierte und hybride Systeme. Dennoch zog das Thema viel Aufmerksamkeit auf sich. Die Entwicklung rein lernbasierter NLP-Komponenten verspricht robuste und schnell entwickelbare Systeme, die den Entwickler vom mühsamen Handwerk des Regelschreibens befreien. Die verfügbaren NER-Datensätze von MUC-6 und MUC-7 boten eine geeignete Aufgabenstellung, um Erfahrungen im Umgang mit lernbasierten Verfahren zu gewinnen. Beides führte dazu, dass NER zu einer bevorzugten Domäne des Maschinellen Lernens für NLP (vgl. dazu die Shared Tasks bei CONLL-2002, CONLL-2003 und JNLPBA-2004) geworden ist und eine Vielzahl von Arbeiten veröffentlicht wurde. Die folgende Darstellung stellt die wichtigsten Aspekte und Beiträge vor, die für das Verständnis der Entwicklung des Gebietes, der im Moment vorherrschenden Ansätze und des in Kapitel 5. vorgestellten Systems bedeutsam sind.

4.2.1. Transformationsbasiertes Regellernen

Alembic ([ABERDEEN ET AL. 1995]) ist das erste System, welches eine Lernerkomponente zur NER integriert. Es basiert auf einer Version des fehlergesteuerten, transformationsbasierten Regellernens, vergleichbar zu Brill's POS-Tagger ([BRILL 1992]). Dabei wird ein Token in einem ersten Schritt aufgrund eines trivialen Algorithmus klassifiziert, z.B. aufgrund der häufigsten NE-Klasse des Tokens in den Trainingsdaten. Diese Zuweisung wird anschließend durch den Einsatz von Regeln verbessert, die aus einer großen Anzahl mittels Transformationen erzeugter Regeln automatisch ausgewählt werden. Die Transformationsregeln beruhen auf einfachen Regelskeletten, beispielsweise der folgenden Form:

> *Ändere die NE-Klasse X, nach NE-Klasse Y, wenn das aktuelle Wort die Wortart Z und das nachfolgende Wort die Wortform W hat.*

Ein Algorithmus probiert alle möglichen bzw. sinnvollen Werte für die Variablen X,Y,Z,W durch, wendet die resultierenden Regeln auf ein Trainingskorpus an und evaluiert jede Regel

anhand der zusätzlichen Treffer bzw. Fehler, die eine Transformation hervorbringt. Transformationsregeln, die zu einer genügend großen Verbesserung führen, werden gespeichert, die übrigen verworfen. Der Komplexität der Regeln sind, insbesondere bei einer Vielzahl möglicher Variablenwerte, aufgrund der erforderlichen Trainingszeit Grenzen gesetzt. Ist das Verfahren einmal trainiert, so besteht es einzig aus einem Algorithmus zur ersten NE-Klassifikation und einer Anzahl einfacher Regeln. In der Anwendung ist es genau so schnell wie ein herkömmliches, regelbasiertes System. Die hohen Trainingszeiten und das Problem der Regelkombination bzw. sich widersprechender Regeln beschränkt allerdings die Möglichkeiten, beliebige Evidenzen in das Modell einfließen zu lassen. [ABERDEEN ET AL. 1995] erzielten enttäuschende Ergebnisse im Vergleich zu den besten MUC-6 Beiträgen. Ob dies an spezifischen Eigenschaften des Ansatzes von [ABERDEEN ET AL. 1995] liegt, lässt sich aufgrund der wenig detaillierten Beschreibung nicht beurteilen. Allerdings sind keine weiteren transformationsbasierten Ansätze zur NER bekannt.

4.2.2. Hidden Markov Modelle

Hidden Markov Modelle (HMM) wurden bereits erfolgreich zur Spracherkennung (vgl. für einen Überblick [RABINER 1989]) und beim POS-Tagging eingesetzt ([CHURCH 1988], [WEISCHEDEL ET AL. 1993]). Mit BBNs Identifinder™ ([BIKEL ET AL. 1997], [MILLER ET AL. 1998], oder ausführlicher [BIKEL ET AL. 1999]) konnte bei der MUC-7 erstmalig gezeigt werden, dass ein rein lernbasierter Ansatz mittels HMM annähernd die Ergebnisse regelbasierter Systeme erzielen kann. HMMs können dazu benutzt werden, einer Sequenz von Beobachtungen eine Sequenz von Tags zuzuordnen. Bei der NER sind die Beobachtungen die Wörter und die Tags die NE-Klassen. Dabei wird angenommen, dass ein stochastischer endlicher Automat, das Markov Modell, die Sequenz von Beobachtungen erzeugt, die Zustände des Automaten aber nicht sichtbar sind. Anhand der Beobachtungen wird nun die wahrscheinlichste Abfolge von Zuständen des endlichen Automaten ermittelt, welcher die beobachtete Sequenz erzeugt hat. Start- und Endzustand ignorierend, kann das Markov Modell auf die Ausgabewahrscheinlichkeit eines Zustandes für eine bestimmte Beobachtung und die Übergangswahrscheinlichkeit zwischen den Zuständen reduziert werden. Dabei wird die sog. Markov Annahme zugrundegelegt, die davon ausgeht, dass der aktuelle Zustand nur von einem (Markov Modell erster Ordnung) oder zwei (Markov Modell zweiter Ordnung) vorherigen Zuständen abhängig ist. Auf eine ganze Sequenz angewandt, bestimmt das HMM

die Wahrscheinlichkeit einer Sequenz von Zuständen Y, gegeben eine Sequenz von Beobachtungen X zu jedem Zeitpunkt T. Für ein Markov Modell erster Ordnung also

$$p(x,y) = \prod_{i=1}^{t} p(y_i \mid y_{i-1}) p(x_i \mid y_i)$$

Daraus kann mittels Viterbi-Dekodierung ([VITERBI 1967]) sehr effizient die wahrscheinlichste Sequenz der Zustände gefunden werden.

Anhand eines annotierten Trainingskorpus werden für die NER die Übergangswahrscheinlichkeiten zwischen den Zuständen, also den NE-Klassen abgeschätzt. Die Wahrscheinlichkeit $p(x \mid y)$, dass ein Zustand ein bestimmtes Wort ausgibt, kann nicht direkt im Korpus beobachtet werden, aber sie kann mit dem Theorem von Bayes aus dem anhand des Korpus abgeschätzten $p(y \mid x)$ bestimmt werden.

[BIKEL ET AL. 1997] beziehen auch deterministische Oberflächenmerkmale des Wortes (vgl. Kap. 3.2.3), das vorangehende Wort und die Wahrscheinlichkeit des Bigramm des aktuellen und des vorangehenden Wortes mit ein. Dies integriert zum einen die, wenn auch sehr lokale, externe Evidenz des vorangehenden Wortes, erlaubt zum anderen auch den erforderlichen Umgang mit unbekannten Wörtern. Aufgrund der korpusbasierten Wahrscheinlichkeitsabschätzung ist die Wahrscheinlichkeit eines nicht im Training gesehenen, also unbekannten Wortes null. Dies ist falsch und führt zum unerwünschten Effekt, dass die multiplikationsbasierte Berechnung der Wahrscheinlichkeit der Sequenz ebenfalls null ergibt. Um dies zu vermeiden, werden ein oder mehrere sogenannte Back-Off Modelle eingesetzt. Diese Modelle sind schwächer, d.h. verfügen über weniger spezifische Informationen, also beispielsweise nur das Oberflächenmerkmal des Wortes, aber nicht das Wort selbst, leiden jedoch weniger unter Ereignissen mit unbekannten Wahrscheinlichkeiten. Ein solches Back-Off Modell kann auch für ungesehene Übergangswahrscheinlichkeiten erforderlich sein, insbesondere wenn nicht wie bei BBNs Identifinder™ nur Bigramm-, sondern auch Trigramm-Wahrscheinlichkeiten benutzt werden.

Je mehr Wahrscheinlichkeiten in das Modell mit einfließen, desto wichtiger sind Verfahren wie beispielsweise die lineare Interpolation, die die unterschiedlichen Wahrscheinlichkeiten gewichten bzw. aufeinander abstimmen und die Wahrscheinlichkeiten gesehener und ungesehener Ereignisse glätten. Die Berücksichtigung weiterer Wahrscheinlichkeiten ist bei

Markov Modellen nur beschränkt möglich. Der generative Ansatz der Markov Modelle versucht die Wahrscheinlichkeit des gemeinsamen Auftretens von Beobachtung und Zustand, also $p(x,y)$ zu generieren bzw. abzuschätzen. Ist die Beobachtung x kein atomares Merkmal, sondern ein komplexer Merkmalsvektor, so wird die Abschätzung des gemeinsamen Auftretens von $p(x,y)$ extrem aufwändig, da die gemeinsame Wahrscheinlichkeit des Auftretens von y für alle Werte und Wertkombinationen des Merkmalsvektors $\vec{x}$ bekannt sein müssen. Da sich in einem annotierten Korpus nur eine überschaubare Anzahl von Wertekombinationen in genügender Zahl beobachten lässt, sind generativen Ansätzen bei der Wahl des Merkmalsvektors, also der Berücksichtigung von Evidenzen, enge Grenzen gesetzt: Der Merkmalsvektor sollte nur Werte zulassen, deren Wahrscheinlichkeit des gemeinsamen Auftretens mit y sich anhand der Trainingsdaten abschätzen lässt.

Eine Alternative zum generativen Vorgehen sind diskriminative Lernverfahren, welche darauf verzichten, die gemeinsame Wahrscheinlichkeit $p(x,y)$ abzuschätzen, sondern sich auf die Abschätzung der bedingten Wahrscheinlichkeit $p(y|x)$ beschränken. Da dadurch nicht mehr die Wahrscheinlichkeit der Beobachtung x benötigt wird, kann der Merkmalsvektor sehr viel komplexer gestaltet werden. Allerdings sind dazu aufwändigere Lernalgorithmen und Optimierungen erforderlich. Prominente Beispiele diskriminativer Verfahren sind Maximum Entropy, Conditional Random Fields oder Support Vektor Maschinen.

In einem gewissen Rahmen können jedoch auch HMMs erweitert werden, so dass sie in der Lage sind, mehrere Merkmale für eine Beobachtung zu verarbeiten. [ZHOU & SU 2002] verwenden nicht nur Trigramme, um einen weiteren lokalen Kontext zu berücksichtigen, sondern auch eine ganze Reihe zusätzlicher Merkmale und übertreffen auf den Daten von MUC-6 und MUC-7 alle bisherigen Systeme. Neben den deterministischen Oberflächenmerkmalen werden wichtige Trigger zu semantischen Kategorien zusammengefasst, das Vorkommen in Namenslisten und ob ein Wort in derselben Texteinheit bereits als NE klassifiziert wurde (vgl. dazu Kapitel 3.2.6). Im Gegensatz zum klassischen HMM optimiert der Ansatz von [ZHOU & SU 2002] nicht die Wahrscheinlichkeit der Sequenz der Beobachtungen, sondern der Tags und addiert logarithmierte Wahrscheinlichkeiten. Darüber hinaus wird ein im Vergleich zu [BIKEL ET AL. 1997] umfangreicheres Tagset benutzt, welches neben der Namensklasse auch angibt, ob ein Wort eine vollständige NE oder

der Beginn, eine Fortsetzung oder das Ende einer NE ist. [ZHOU & SU 2002] beschreiben mehrere Back-Off Modelle, welche unterschiedlichste Merkmalsklassen kombinieren, beschreiben jedoch nicht, nach welcher Strategie diese gewählt und gewichtet werden. Allerdings wird in [ZHOU & SU 2003] ein Algorithmus vorgestellt, der iterativ nach der besten Merkmalskombination sucht.

4.2.3. Maximum Entropy

Maxim Entropy (ME) gehört zu den diskriminativen Ansätzen und unterliegt damit nicht den in 4.2.2 beschriebenen Beschränkungen der generativen HMMs. Im Gegensatz zum generativen Ansatz modellieren diskriminative Ansätze nicht $p(x,y)$, also die Wahrscheinlichkeit des gemeinsamen Auftretens einer Beobachtung und einer Klassifikation, sondern sie beschränken sich auf $p(y|x)$, für NER also auf die bedingte Wahrscheinlichkeit des Auftretens einer NE-Klasse y, gegeben ein Wortvorkommen x. Damit wird nicht mehr die Wahrscheinlichkeit $p(x)$ einer Beobachtung benötigt, welche sich im Falle eines komplexen Merkmalsvektors $\vec{x}$ kaum mehr zuverlässig aus einem annotierten Korpus abschätzen lässt. Dies macht den Weg frei für Verfahren, welche die konditionale Wahrscheinlichkeit $p(y|x)$ für einen komplexen Merkmalsvektor $\vec{x}$ modellieren. Die Vielzahl an Evidenzen, die zur NER zur Verfügung stehen, können so bei der Modellierung berücksichtigt werden.

Maximum Entropy Modellierung ist eines der ersten diskriminativen Lernverfahren, welches für die NER eingesetzt wurde, und stand bereits zu MUC-7 Zeiten als fertig implementierte Lösung zur Verfügung (z.B. [RISTAD 1998]). Einen Überblick über weitere NLP-Anwendungen mit ME bietet [BERGER ET AL. 1996].

Ein ME-Klassifizierer berechnet für ein Wortvorkommen x die Wahrscheinlichkeiten der jeweiligen Klassenzugehörigkeiten y. ME erlaubt die Berücksichtigung einer Vielzahl von binären Merkmalen, welche durch die Merkmalsfunktionen $F(X)$ erzeugt werden, und berechnet für jedes Merkmal $f(x,y)$ ein Gewicht λ:

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n} \lambda_i f_i(x,y) \right)$$

wobei $Z(x)$ nur ein Normalisierungsfaktor ist.

Bei der Suche nach den Gewichten der einzelnen Merkmale wird nach dem Prinzip der maximalen Entropie vorgegangen. Gesucht wird das Modell, welches außer den Beobachtungen in den Trainingsdaten keine Voraussagen macht. Das bedeutet, das Modell, welches das aus den Trainingsdaten abgeleitete Wissen berücksichtigt, aber darüber hinaus die größtmögliche Ungewissheit, also die maximale Entropie annimmt. Entropie ist das anerkannte Maß für die Gewissheit bzw. Ungewissheit von Informationen und stammt aus der Informationstheorie ([SHANNON 1948]).

Um das Modell mit der maximalen Entropie zu finden, können verschiedene iterative Algorithmen verwendet werden (vgl. [MALOUF 2002] für eine Evaluation der verschiedenen Algorithmen). Insbesondere bei der Gewichtung seltener Merkmale neigen ME Modelle zum Overfitting. Diesem kann mit Verfahren zur Glättung der Gewichte begegnet werden (vgl. beispielsweise [CHEN & ROSENFELD 1999]). Wird ME als Tagger verwendet, so wird aus den berechneten Wahrscheinlichkeiten der Klassenzugehörigkeit meist mittels Viterbi-Dekodierung ([VITERBI 1967]) die wahrscheinlichste Sequenz ermittelt.

Das System MENE (Maximum Entropy Named Entity, [BORTHWICK ET AL. 1998], [BORTHWICK 1999]) erzielte bei der MUC-7 als zweites, vorrangig lern-basiertes System vielbeachtete, wenn auch schwächere Ergebnisse als BBNs Identifinder™ ([BIKEL ET AL. 1997]). MENE wird hier ausführlicher als andere Ansätze besprochen, da es prototypisch für NER-Systeme ist, welche ein Wortvorkommen mit einer Vielzahl von Merkmalen repräsentieren und es einem ML-Verfahren überlassen, daraus ein Modell zur NER abzuleiten.

MENE weist jedem Wort eine Auszeichnung zu, welche die NE-Klasse anzeigt und zusätzlich, ob das einzelne Token ein vollständiger Name ist, eine NE-Sequenz startet, weiterführt oder beendet. Dies ergibt vier Label pro NE-Klasse und eine zusätzliche Klasse für Nicht-NEs. MENE benutzt eine Vielzahl von Evidenzen, die alle für die ME Modellierung auf binäre Merkmale abgebildet werden müssen.

Für ein Fenster von fünf Wörtern, nämlich zwei vor und zwei hinter dem zu klassifizierenden Wort, werden folgende Merkmale benutzt:

- Deterministisches Wortoberflächenmerkmal
- Lexikalisches Merkmal: Ist ein Wort mindestens dreimal in den Trainingsdaten gesehen worden, so gilt es als bekannt und das Vorkommen der Wortform an einer

bestimmten Stelle wird als Merkmal benutzt. Ansonsten wird das Merkmal „unbekanntes Wort“ erzeugt.

- Dokumentenstruktur: Kommt ein Wort in einer Überschrift vor, so wird dies als Merkmal angegeben. Das Merkmal soll abbilden, dass Wörter in Überschriften öfter NEs enthalten als Wörter im Hauptteil eines Textes. Allerdings zeigte das Merkmal in den Experimenten keinen Einfluss auf die Performanz.
- Lexikalische Ressourcen: MENE hat Zugriff auf umfangreiche, teils semi-automatisch, teils manuell erstellte Listen von Personen-, Orts-, und Organisationsnamen. Zu allen darin vorkommenden Wörtern wird neben der Namenskategorie vermerkt, ob sie am Beginn, innerhalb, am Ende eines Namens stehen oder auch als einzelnes Wort einen ganzen Namen bilden.
- Die Voraussage eines regelbasierten Systems: Für die MUC-Evaluation hat MENE Zugriff auf die Ausgaben des regelbasierten Systems von [GRISHMAN 1995], welche ebenfalls auf Merkmale abgebildet werden. Dadurch kann MENE den Voraussagen des regelbasierten Systems widersprechen und wird auch in der Hoffnung eingesetzt, systematische Fehler des regelbasierten Systems zu erkennen und zu korrigieren. Die Ausgabe des regelbasierten Systems hat großen Einfluss auf die Leistung von MENE und verbessert das F-Measure um fast fünf Punkte.
- In späteren Versionen von MENE (beschrieben in [BORTHWICK 1999]) wird eine zweistufige Klassifikation verwendet. Die zweite Stufe hat dabei Zugriff auf die Ausgabe der ersten und insbesondere darauf, ob ein Wort an anderer Stelle im selben Text als NE markiert wurde.

Mit einer gänzlich anderen Architektur bzw. einer gänzlich anderen Integration der ME-Komponente bewiesen [MIKHEEV ET AL. 1998] den Vorteil einer Lernkomponente in Verbindung mit einem regelbasierten System. Bei MUC-7 wurden mit einem Gesamt-F-Measure von 93.39 alle anderen Systeme übertroffen. Die NER besteht aus fünf Stufen, an denen zweimal ME zur Klassifikation von NE-Kandidaten benutzt wird, die aufgrund von unsicheren Regeln vorgeschlagen werden. Die erste Stufe besteht nur aus so genannten „Sure-fire“-Regeln, welche sowohl aufgrund interner als auch externer Evidenz mit hoher Precision, aber geringem Recall NEs erkennen. Die dabei gefundenen NEs werden zur Erzeugung neuer Kandidaten genutzt. Wird beispielsweise „*Adam Kluver Ltd*“ gefunden, so wird nach Vorkommen von „*Adam Kluver*“, „*Kluver Ltd*“ und „*Adam Ltd*“ gesucht. Diese werden jedoch nicht als NE ausgezeichnet, sondern erst aufgrund eines ME Modells klassifiziert.

Dabei berücksichtigt das Modell unter anderem die in Schritt 1 gefundenen NEs, die Position im Satz und andere, möglicherweise kleingeschriebene Vorkommen derselben Wörter. Auf der dritten Stufe werden, ohne den Kontext zu beachten, aufgrund der bisher gefundenen NEs, aber auch mit Hilfe der NE-Listen, weitere NEs klassifiziert. Dass dabei nicht sehr viele mehrdeutige Wortformen fehlerhaft klassifiziert werden, wird folgendermaßen begründet: Wörter, die abweichend von ihrer üblichen Bedeutung benutzt werden (z.B. „*Washington*“ als Person, Vornamen-Nachnamen-Kombinationen als Organisationsname), würden mit einem Kontext eingeführt, der nicht nur dem menschlichen Leser, sondern auch den Stufen 1 und 2 die NER ermöglichte. Sind solche Vorkommen erkannt, kann unter Vernachlässigung des Kontextes aufgrund von Lexikoneinträgen klassifiziert werden. In Stufe 4 wird erneut aufgrund der bereits in der Texteinheit erkannten Namen klassifiziert. Dabei werden alle möglichen Bestandteile von Namen, die isoliert im Text vorkommen, als NE-Kandidaten betrachtet und aufgrund eines weiteren ME Modells klassifiziert. Die NER wird in Stufe 5 durch die Annotation der durchgängig in Großbuchstaben gesetzten Überschrift aufgrund der bereits gefundenen NEs abgeschlossen.

4.2.4. Conditional Random Fields

Ein sehr interessanter Ansatz sind Conditional Random Fields (CRF), eingeführt von [LAFFERTY ET AL. 2001], welche als Erweiterung der HMMs und der Modellierung mit ME betrachtet werden können. CRF erweitern HMMs, da sie genau wie ME diskriminativ und nicht generativ trainiert werden. Darüber hinaus modellieren CRF im Gegensatz zu ME die gesamte Sequenz und nicht nur den aktuellen und den Vorgängerzustand, wie dies bei Maximum Entropie Markov Modellierung (MEMM) der Fall ist. MEMM können Opfer des sog. Label Bias werden [LAFFERTY ET AL. 2001]. Der Label Bias beschreibt, dass Zustände mit wenigen Nachfolgern automatisch eine höhere Wahrscheinlichkeit haben. CRFs überwinden diese Schwierigkeit, indem sie die Möglichkeit bieten, alle oder einige Zustände miteinander zu verknüpfen.

CRF modellieren nicht das wahrscheinlichste Label einer Beobachtung zu einem Zeitpunkt, sondern die wahrscheinlichste Sequenz von Labeln oder Zuständen y, gegeben eine Sequenz von Beobachtungen x. Die Wahrscheinlichkeit der Sequenz wird dabei folgendermaßen definiert:

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(, y_i, x, i) \right)$$

Wobei $Z(x)$ ein Normalisierungsfaktor ist, $t(y_{i-1}, y_i, x, i)$ Zustandsübergangsmerkmale und $s(y_i, x, i)$ Zustandsmerkmale jeweils an der i-ten Stelle der Sequenz x sind. λ und μ sind Parameter, welche aufgrund der Trainingsdaten geschätzt werden. Die Zustandsübergangsmerkmale basieren auf der Markov Annahme und erlauben es, für einen bestimmten Zustandsübergang vom Vorgänger zum aktuellen Zustand für eine Beobachtung eine Merkmalsfunktion zu definieren.

$$f_m(y_{i-1}, y_i, x, i) = \left\{ \begin{array}{l} 1 \text{ wenn} \\ (\text{nur\_aus\_Großbuchstaben}(x_i)) = \text{true} \ \& \\ y_{i-1} = \text{ORG} \ \& \ y_i = \text{ORG} \\ 0 \text{ sonst} \end{array} \right\}$$

Die Merkmalsfunktion f_m etwa ergibt 1, wenn der aktuelle und der Vorgängerzustand ORG ist und das aktuelle Wort nur aus Großbuchstaben besteht. Die in einem Merkmal berücksichtigten Zustände bzw. Labels sind auf die aktuelle und die Vorgängerposition beschränkt, während die Beobachtung von x_i nahezu ohne Einschränkung ist, also beispielsweise auch die Beobachtung „an der vierten Position nach dem aktuellen Wort befindet sich ein Punkt“ miteinbeziehen könnte. Die Zustandsmerkmale sind analog zu verstehen, berücksichtigen jedoch nur den aktuellen Zustand.

CRF sind attraktiv, da sie im Gegensatz zu HMMs problemlos mit einer Vielzahl von Merkmalen umgehen können und dennoch ideal sind, um Sequenzen und nicht nur einzelne Tokens zu klassifizieren. Den Vorteil dieser Eigenschaften für das Klassifizieren sprachlicher Sequenzen zeigten [SHA & PEREIRA 2003] auf den Daten des Shared Task für Text-Chunking der CoNLL-2000 ([TJONG KIM SANG & BUCHHOLZ 2000]), wobei sie die besten jemals erreichten Resultate erzielten.

In [MCCALLUM 2003] wird ein verbesserter Trainingsalgorithmus für CRF vorgestellt, der nicht nur effizienter ist, sondern auch aus zufälligen Kombinationen existierender Merkmale weitere Merkmale erzeugt und damit bessere Resultate erreicht, als das in [LAFFERTY ET AL. 2001] beschriebene Verfahren. In [MCCALLUM & LI 2003] wurden CRF erstmalig auf NER

angewandt. Allerdings wurden nur durchschnittliche Resultate erzeugt. [SETTLES 2004] hingegen demonstrierte mit seinem Beitrag zum Shared Task der JNLPBA-2004 den erfolgreichen Einsatz bei der NER in der biomedizinischen Domäne.

4.2.5. Support Vektor Maschinen

Die Stützvektormethode (Support Vector Machine, SVM), eingeführt von [VAPNIK 1995], ist ein diskriminativer Lernalgorithmus der mit sehr großen Merkmalsmengen umgehen kann, robust gegen Overfitting ist und bereits mehrfach erfolgreich für die Verarbeitung natürlicher Sprache eingesetzt wurde (Textkategorisierung [JOACHIMS 1998], Text-Chunking [KUDOH & MATSUMOTO 2000], POS-Tagging [GIMÉNEZ & MÀRQUEZ 2003]). Im Gegensatz zu anderen diskriminativen Lernverfahren modelliert die SVM nicht die konditionale Wahrscheinlichkeit der Klassenzugehörigkeit einer Beobachtung, sondern beschränkt sich auf die Suche nach dem optimalen Klassifizierer.

Da die SVM auch für das im Rahmen des Dissertationsprojekts entwickelte System eingesetzt wird, wird das Verfahren ausführlicher als die anderen vorgestellt. Wie in den folgenden Ausführungen deutlich werden wird, ist eine SVM immer auf eine binäre Klassifikation beschränkt, so dass für NER mehrere SVMs erforderlich sind. Die Kombination der Ergebnisse der SVMs wird weiter unten besprochen. Für die folgende Einführung wird diese Problematik vorerst ausgeklammert. Es wird also immer nur eine binäre SVM-Aufgabe betrachtet, also ob es sich beispielsweise um eine NE vom Typ Organisation handelt oder nicht.

Die SVM stammt aus der statistischen Lerntheorie und beruht auf der Idee der strukturellen Risikominimierung. Für eine detailliertere Beschreibung der SVM und der zugrunde liegenden Theorie sei auf [VAPNIK 1995], das Tutorial von [BURGES 1998] oder die Einführung von [CHRISTIANINI & SHAWE-TAYLOR 2000] verwiesen. Die folgenden Ausführungen orientieren sich an [JOACHIMS 2002].

Eine SVM erhält als Eingabe eine Anzahl von Trainingsbeispielen der folgenden Form:

$$(\vec{x}_1, y_1),...,(\vec{x}_l, y_l) \text{ mit } (\vec{x}_i, y_i) \in \Re^n \times \{+1, -1\} \quad (1)$$

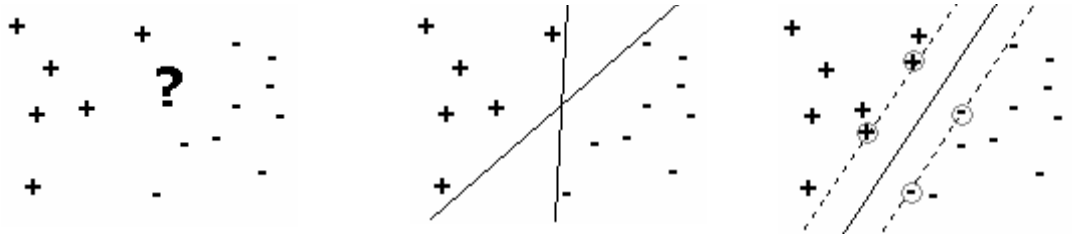
Ein einzelnes Beispiel $(\vec{x}_i, y_i)$ ist durch den n-dimensionalen Merkmalsvektor $\vec{x}_i$ und das Klassenlabel y_i beschrieben. Gesucht ist die Klassifikationsregel, welche neue Beispiele vom selben Typ möglichst korrekt klassifiziert. Eine Klassifikationsregel h kann folgendermaßen dargestellt werden:

$$h(\vec{x}) = sign\left(b + \sum_{i=1}^{n} (w_i \cdot x_i) \right) = sign(\vec{w} \cdot \vec{x} + b) \quad (2)$$

Im Gegensatz zu anderen diskriminativen Verfahren wie den Conditional Random Fields oder Maximum Entropy wird also nicht die konditionale Wahrscheinlichkeit $p(y|x)$ modelliert; die SVM beschränkt sich auf eine Funktion bzw. Klassifikationsregel $h(\vec{x})$, welche das Ergebnis einer binären Klassifikationsaufgabe durch die Ausgabe von -1 für die eine, und +1 für die andere Klasse anzeigt.

Die Klassifikationsregel besteht aus dem Gewichtsvektor $\vec{w}$, der jedem Merkmal ein Gewicht zuordnet, und dem Schwellwert b. Im Training legt die SVM sowohl $\vec{w}$ als auch b fest. Ein unbekanntes Beispiel mit zugehörigem Vektor $\vec{x}$ wird klassifiziert, indem $h(\vec{x}) = sign(\vec{w} \cdot \vec{x} + b) = y$ berechnet wird: Also die Bildung des Skalarprodukts $\vec{w} \cdot \vec{x}$, die Addition von b und der Anwendung der Funktion $sign: \Re \rightarrow$ {+1, -1}.

Durch den Ausdruck $sign(\vec{w} \cdot \vec{x} + b)$ wird eine mehrdimensionale Ebene, eine so genannte Hyperebene beschrieben. Diese trennt die positiven von den negativen Beispielen. Für den zweidimensionalen Raum genügt eine Gerade. Wie in Abbildung 4.1 gezeigt wird, kann eine Menge von positiven und negativen Beispielen durch unterschiedliche Geraden getrennt werden.



Aufgabe | **Trennmöglichkeiten** | **Trennebene mit größtem Abstand und eingekreisten Support Vektoren**

Abbildung 4.1: Binäre Klassifikationsaufgabe

Die SVM wählt die Hyperebene, welche die positiven und negativen Beispiele mit maximalem Abstand trennt. Die Annahme der SVM ist also, dass die Breite des Randes ein

Maß für die Generalisierung darstellt, also der Güte für zukünftige Trennaufgaben. Die Instanzen, die der trennenden Ebene am nächsten liegen, werden Support Vektoren genannt.
Das Auffinden der Lösung mit dem maximalen Abstand wird als quadratisches Optimierungsproblem formuliert:

$$\min_{(\alpha_1,\ldots,\alpha_l)} -\sum_{i=1}^{l}\alpha_i + \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\vec{x}_i \cdot \vec{x}_j) \tag{3}$$

so dass $\alpha_i \geq 0$ für $i = 1,\ldots,l$ (4)

und $\sum_{i=1}^{l} y_i \alpha_i = 0$ (5)

Gesucht werden also die so genannten Lagrange-Multiplikatoren $\alpha_1 \ldots \alpha_l$, die alle größer oder gleich null sind und für die auch (5) gilt. Nach der Berechnung der Lösung $\alpha_1^* \cdots \alpha_l^*$ ergibt sich folgende Klassifikationsregel:

$$h(x) = sign\left(b + \sum_{i=1}^{l} (\alpha_i^* y_i \vec{x}_i \cdot \vec{x}) \right) \tag{6}$$

Der Gewichtsvektor $\vec{w}$ wird berechnet durch

$$\vec{w} = \sum_{i=1}^{l} \alpha_i^* y_i \vec{x}_i \tag{7}$$

wobei nur die $\alpha_i > 0$ in die Berechnung eingehen. Damit stellt die Klassifikationsregel eine Kombination von Lernbeispielen dar. Der Schwellwert lässt sich mit jedem beliebigen Stützvektor durch Einsetzen in $b = y_i - \vec{w} \cdot \vec{x}_i$ bestimmen.

Manche Datensätze können nicht linear getrennt werden. Dann setzt die SVM die so genannte weiche Trennung ein, nimmt also eine gewisse Fehlklassifikation der Trainingsinstanzen in Kauf. Die Aufgabe ist also nicht nur, die optimale Trennebene zu finden, sondern auch den Fehler εi zu minimieren. Dabei misst εi, wie weit im oder jenseits des Trennbereichs ein Beispiel liegt.

Berechne $\vec{w}, b$ (8)

so dass $w \cdot w + C \sum_{i=1}^{l} \varepsilon_i$ minimal (9)

und es gilt $y_1[\vec{w}\cdot\vec{x}_1+b]\geq 1-\varepsilon_1,\ldots,y_l[\vec{w}\cdot\vec{x}_l+b]\geq 1-\varepsilon_l$ (10)

Der Parameter C erlaubt das Abwägen von Trennabstand und Fehler. Das duale quadratische Optimierungsproblem entspricht dem oben ausgeführten, wenn (4) erweitert wird durch

$$0\leq\alpha_i\leq C \text{ für } i=1,\ldots,l \quad (11)$$

Nicht-lineare SVMs bilden den Merkmalsraum mittels Funktionen auf einen höherdimensionalen Raum ab. Dies wird als so genannter Kernel-Trick bezeichnet. Man bildet dazu beispielsweise aus zwei Merkmalen ein neues Merkmal, das aussagt, ob die beiden Merkmale gemeinsam auftreten, oder ob nur eines der beiden Merkmale auftritt. Diese neue Abbildung ist oft leichter zu trennen, das heißt, sie führt oft zu besseren Ergebnissen in der Klassifikation. Werden Merkmale als Zahlen kodiert, lassen sich neue Merkmale etwa als Produkt von existierenden Merkmalen erzeugen. Eine ausführlichere Darstellung nicht-linearer SVMs findet sich beispielsweise in [CHRISTIANINI & SHAWE-TAYLOR 2000].

Wie sich aus den Ausführungen ergibt, verhält sich der Trainingsaufwand der SVM quadratisch zur Anzahl der Instanzen und nicht zur Anzahl der Merkmale bzw. Dimensionen. Das ist zwar einerseits wünschenswert, da damit die Grundlage zur Verarbeitung von hochdimensionalen Klassifikationsaufgaben gegeben ist, so dass es möglich ist, nahezu beliebige Evidenzen im Modell zu berücksichtigen. Andererseits erfordert es, die Anzahl der Instanzen in einem gewissen Rahmen zu halten, damit der Trainingsaufwand nicht ins Unermessliche steigt. Damit ist die SVM eher ungeeignet für Verfahren, welche die Trainingsinstanzen durch ungelabelte Daten expandieren. Derartige Erweiterungen der Trainingsinstanzen werden auch für NER eingesetzt und werden in Abschnitt 4.3 vorgestellt. Die Anzahl der Instanzen ist bei der NER aufgrund der tokenbasierten Herangehensweise nicht gering. Bereits das CoNLL-Korpus enthält etwa 220.000 Wörter, von denen je nach Ansatz zwischen 100.000 und 150.000 als Instanzen herangezogen werden, was zu einem gewissen Trainingsaufwand führt, der jedoch für lineare SVMs noch keinerlei Schwierigkeiten bereitet. Für größere Trainingskorpora jedoch, wie beispielsweise die 500.000 Wörter mit biomedizinischen Entities ([JNLPBA-2004]), lässt der Trainingsaufwand nur noch lineare SVMs zu. [TAKEUCHI & COLLIER 2002] etwa führen ihre Experimente mit polynomialen Kernfunktionen aufgrund des Trainingsaufwandes nur auf einem Ausschnitt der zur Verfügung stehenden annotierten Daten durch.

Die Beschränkung auf binäre Klassifikationen erfordert eine Binarisierung der NE-Klassifikation. Hierzu sind mehrere Kombinationen von Klassifikationsaufgaben denkbar (vgl. dazu [HSU & LIN 2001]). Allerdings scheint sich beim erfolgreichen Einsatz der SVM in der NER die einfache „ein-Klassifizierer-pro-NE-Klasse“-Strategie ([MAYFIELD ET AL. 2003], [ZHOU & SU 2004]) durchzusetzen. Eine der Aufgabe angemessen scheinende Aufteilung, wie beispielsweise das Erkennen von Namen in einem ersten Schritt und die Klassifikation derselben in einem zweiten Schritt ([LEE ET AL. 2003]), hat sich bisher nicht bewährt.

Der Einsatz von SVMs beschränkt sich auf Bereiche, die schwieriger als die klassische MUC-NER sind und die Berücksichtigung einer Vielzahl von Merkmalen zu erfordern scheinen. Neben den bereits für englische NER eingesetzten Merkmalen wie Wortform, Wortart, Wortoberflächenmerkmale sind dies v.a. große Listen von Substrings aber auch eine Vielzahl von Namenslisten unterschiedlichster Herkunft. [MAYFIELD ET AL. 2003] und [RÖSSLER 2004a] setzen sie für NER in deutschen Texten ein. Für NER in biomedizinischen Texten wurde die SVM erstmals von [TAKEUCHI & COLLIER 2002] angewendet und war am Shared Task der JNLPBA-2004 ([KIM ET AL. 2004]) das am meisten eingesetzte Verfahren. Ansätze, die sich jedoch vorrangig auf Klassifikationseigenschaften der SVM verließen, wurden am Shared Task der [JNLPBA-2004] durch Verfahren übertroffen, welche die SVM mit HMMs verbanden ([ZHOU & SU 2004]), durch ME-Modellierung ([FINKEL ET AL. 2004]), aber auch durch Conditional Random Fields ([SETTLES 2004]). Der Verdacht liegt nahe, dass die mehrfach bewiesene Fähigkeit der SVM mit großen Merkmalsräumen umzugehen, für die Erkennung langer und komplexer NEs nicht genügt, sondern einer stärkeren Berücksichtigung der Namen als eine Sequenz von Wörtern bzw. Tokens bedarf. Dieser Aspekt wird in den überlegeneren Verfahren meist mit Hilfe der Viterbi- Dekodierung modelliert oder ist, wie bei CRF, integraler Bestandteil der Lernaufgabe. CRF modellieren eine Sequenz von NE-Labeln y, gegeben eine Sequenz von Wörtern x, während mit SVMs, genau wie bei ME, ein NE-Label y zu einem Wort x gesucht wird. Die Berücksichtigung der Tatsache, dass viele NEs aus mehreren Wörtern bestehen, oder, allgemeiner, der Linearität von Sprache kann bei der wortorientierten Herangehensweise der SVM durch die N-Gram Modellierung des Kontexts, also der Berücksichtigung der Wörter vor und hinter dem zu klassifizierenden Wort und einer anschließenden Zusammenführung bzw. Linearisierung der Klassifikationsergebnisse erreicht werden. In Abschnitt 5.2.1 wird darüber hinaus eine erweiterte N-Gram Modellierung vorgestellt, welche den Mehrwort-Charakter der NEs angemessener erfasst. Wünschenswert

ist jedoch eine Berücksichtigung nicht nur der umgebenden Wörter, sondern auch der vorangegangenen Klassifikationsentscheidung. In [ALTUN ET AL. 2003] wird eine Erweiterung der SVM mit der Markov Modellierung in Form einer Hidden Markov Support Vektor Maschine vorgestellt. Dabei wird die SVM zur Klassifikation von Sequenzen optimiert, indem Beobachtungen nicht nur auf das aktuelle, sondern auch auf das Vorgängerlabel bezogen werden. Leider stand zum Zeitpunkt des Dissertationsprojekts keine effiziente Implementation des Verfahrens zur Verfügung, so dass auf den Einsatz dieser für NER äußerst interessanten Erweiterung verzichtet werden musste.

4.2.6. Weitere eingesetzte ML-Verfahren

Entscheidungsbäume ([QUINLAN 1993]) gehören zu den frühesten für die NER eingesetzten Lernverfahren ([BENNETT ET AL. 1997], [SEKINE 1998]), können jedoch schlecht mit großen Merkmalsmengen umgehen, also beispielsweise die Berücksichtigung der einzelnen Wortformen, und wurden nachfolgend kaum mehr eingesetzt. Memory-based Learning, also ähnlichkeitsbasiertes Lernen wie beispielsweise in TiMBL ([DAELEMANS ET AL. 2003]), ist ein intuitives Lernverfahren, welches mit einer großen Anzahl von Merkmalen umgehen kann. Allerdings schnitten darauf aufbauende Ansätze bei den NER Evaluationskampagnen CONLL-2002 ([TJONG KIM SANG 2002a]) und CONLL-2003 ([DE MEULDER & DAELEMANS 2003]) unterdurchschnittlich schlecht ab. Möglicherweise liegt dies an TiMBLs Schwierigkeiten im Umgang mit redundanten Merkmalen. Eine solche Eigenschaft stellt für natürlichsprachliche Daten, die viele Redundanzen enthalten, einen ernsthaften Nachteil dar.

4.3. Der Einsatz nicht-annotierter Daten

Während das manuelle Annotieren von Daten, das Schreiben von Regeln und das Sammeln bzw. Filtern von Namenslisten zeitaufwendig und damit teuer ist, stehen nicht-annotierte Daten meist preiswert und in großen Mengen zur Verfügung. Diese für die NER nutzbar zu machen, bietet sich deswegen sowohl für regelbasierte als auch für ML-basierte Systeme an. Der Vorteil der preiswerten Verfügbarkeit paart sich im Idealfall mit der Domänen-Relevanz des extrahierten Wissens. Schließlich besteht die Attraktivität korpusbasierter Ansätze darin, dass sie nicht Gefahr laufen, Phänomene zu modellieren, die in den schlussendlich zu verarbeitenden Daten gar nicht vorkommen. Der selbst entwickelte Ansatz zum Lernen aus nicht-annotierten Daten ist in [RÖSSLER 2004b] publiziert. Er steht zwar in Verbindung mit den hier vorgestellten Verfahren, beschreitet jedoch eine ganz andere Richtung. Während die

hier vorgestellten Systeme nicht annotierte Daten zum Erzeugung neuer Trainingsinstanzen oder neuer Namens- oder Kontextslisten einsetzen, verfolgt der eigene Ansatz die Anreicherung der existierenden Trainingsinstanzen durch Merkmale, welche aufgrund automatisch annotierter Daten gewonnen werden. Er wird in den Kapiteln 5. und 6. ausführlicher vorgestellt.

Bisherige Ansätze zum Einsatz nicht-annotierter Daten

Davon ausgehend, dass eine Namensdatenbank niemals vollständig sein wird, wurde bereits 1993 von [MANI ET AL. 1996] ein System vorgeschlagen, welches in Texten selbständig Namen sammelt und diesen aufgrund des Kontextes semantische Eigenschaften, vergleichbar der NE-Klassifizierung, zuweist. Außerordentlich fortschrittlich, wenn auch weitgehend unbeachtet, ist die Arbeit von [THIELEN 1995] über deutsche Eigennamen. Mit der Absicht, POS-Taggern die Unterscheidung von gewöhnlichen Nomen und Eigennamen zu erleichtern, hat sie ein System entwickelt, welches aufgrund eines initialen Sets von Wortbestandteilen, Kontexten oder der Eigenschaft der Artikellosigkeit Eigennamenkandidaten sucht. Die möglichen Eigennamen werden aufgrund von einfachen Regeln gefiltert. Beispielsweise werden großgeschriebene Einheiten dann als mögliche Eigennamen zurückgewiesen, wenn sie vor einem Genitiv-Artikel oder einem Partizip Perfekt stehen, da damit laut [THIELEN 1995] Redewendungen erkannt und ausgeschlossen werden, wie beispielsweise „*aus Anlass des*“ oder „*in Kauf genommen*“. Die verbleibenden Einträge werden aufgrund von Korpusvorkommen gefiltert, und zwar danach, ob sie ohne einleitenden Artikel vorkommen. Dadurch werden jedoch Institutionsnamen („*die CDU*“) und einige Ortsnamen („*die Schweiz*“) ausgeschlossen. Iterativ werden die als gültig gewerteten Namen dazu benutzt, neue Kontexte zu extrahieren, die erneut unbekannte Namen entdecken.

Diese Ausnutzung der Komplementarität externer und interner Evidenz, ausgehend von einem initialen NER-Modell, ist die Grundlage aller Verfahren zum Einsatz nicht-annotierter Daten, die in späteren Arbeiten als Bootstrapping, halb- oder unüberwachtes Lernen bezeichnet werden. Die Ansätze können nach den folgenden Punkten unterschieden werden:

- Die Repräsentation der linguistischen Einheiten: Hierbei stehen dieselben Möglichkeiten wie bei der Entwicklung herkömmlicher NER-Verfahren zur Verfügung (vgl. dazu Kap. 3.2).

- Das initiale NER-Modell: Dieses kann recht einfach aus wenigen, aber verlässlichen Regeln oder einer Anzahl von NEs bestehen, aber es kann auch ein komplexeres, Regel- oder ML-basiertes System verwendet werden.
- Das Extraktionsziel: Es können Namensbestandteile, ganze Namen und/oder Kontexte extrahiert werden. Auch könnte gezielt oder implizit nach Namen und Namensbestandteilen gesucht werden, die in Bezug auf ihre NE-Klasse mehrdeutig sind. Allerdings gehen alle hier besprochenen Ansätze implizit oder explizit ([CUCCHIARELLI & VELARDI 1999], [RILOFF & JONES 1999]) von einer eindeutigen NE-Klasse pro Korpus aus.
- Die Bewertung bzw. Auswahl des extrahierten Wissens: Um zu verhindern, dass falsche Namen und Kontexte in die Bootstrapping-Schleife gelangen, kann das extrahierte Wissen nach Verlässlichkeit sortiert oder ausgewählt werden.
- Das Stopp-Kriterium: Die Bootstrapping-Schleife kann grundsätzlich endlos wiederholt werden, doch wird sie ab einem bestimmten Zeitpunkt nur noch falsches oder gar kein neues Wissen mehr extrahieren. Deswegen wird der Vorgang nach bestimmten Kriterien beendet.

Der Ansatz von [CUCCHIARELLI & VELARDI 1999] zielt auf die Selbstanpassung von Ressourcen für die Informationsextraktion. Als initiales NER-Modell werden Regeln und Namenslisten von nicht beschriebenem Umfang eingesetzt und auf italienische und englische Texte angewendet. Die Umgebung der gefundenen Namen wird mit Hilfe eines Shallow-Parsers analysiert, um daraus korpustypische Kontexte zu extrahieren. Diese werden in das NER-Modell integriert, um iterativ Namen zu lernen und diese den Namenslisten hinzuzufügen. Neue Namen werden aufgrund eines Ähnlichkeitsmaßes, basierend auf syntaktischen und semantischen Eigenschaften aller Vorkommen im Korpus, angenommen oder verworfen. Nach jedem Durchgang wird das Modell auf einem Testkorpus evaluiert. Der Vorgang wird abgebrochen, wenn sich die Performanz nicht mehr verbessert. In den Experimenten war dies nach der dritten Runde der Fall.

Von großer Bedeutung sind die Arbeiten von [RILOFF & JONES 1999] und [THELEN & RILOFF 2002]. Der Ansatz zielt nicht explizit auf die Extraktion von NEs, sondern erzeugt automatisch semantische Lexika von NPs für die Informationsextraktion. Zentral ist das wechselseitige Bootstrapping von Kontexten und den zu den gesuchten Klassen gehörenden NPs. Als Kontexte werden nur bestimmte Syntaxkonstruktionen betrachtet, welche von AutoSlog ([RILOFF 1993], [RILOFF 1996]) extrahiert werden. Die zulässigen

Syntaxkonstruktionen, Extraktionsmuster genannt, sind beispielsweise „<subject> passive verb“ (Beispielsinstanz: „<victim> was murdered“) oder „verb infinitive <direct-object>“ (Beispielsinstanz: „threatened to attack <target>) und werden aufgrund von vorgegebenen POS-Mustern ausgewählt. Alle von AutoSlog in einem nicht-annotierten Korpus extrahierten Extraktionsmuster werden gespeichert und ausgehend von einem ausgesuchten Set von NPs bewertet, die verlässlich zur gesuchten semantischen Klasse gehören. Diese Wertung berücksichtigt, wie oft ein Extraktionsmuster die semantische Kategorie voraussagt und wie viele unterschiedliche NPs der semantischen Klasse gefunden werden. Von allen Extraktionsmustern werden nur die am besten bewerteten benutzt, um neue NPs der gesuchten semantischen Klasse zu extrahieren. Auch die gefundenen NPs werden bewertet, wobei diejenigen höher gewichtet werden, die von mehreren Kontexten vorausgesagt werden. Die am höchsten bewerteten NPs werden dem initialen Set von NPs der gesuchten semantischen Klasse hinzugefügt. Das Verfahren endet üblicherweise nach 50 Durchgängen. Allerdings kann die Wiederholung früher abgebrochen werden, wenn die Bewertung des besten neuen Extraktionsmusters unter einen gewissen Wert fällt, oder fortgesetzt werden, wenn das beste neue Extraktionsmuster über einem gewissen Wert liegt. Als Resultat des Verfahrens kann sowohl auf die NPs der gesuchten semantischen Klassen als auch auf die gefundenen Extraktionsmuster zugegriffen werden.

[COLLINS & SINGER 1999] präsentieren mehrere Ansätze zum Einsatz nicht annotierter Daten für die NER. Für alle wird ein äußerst kleines Set von initialem Wissen zur Verfügung gestellt. Dieses initale Set enthält folgende Angaben: „*New York*“, „*California*“ und „*U.S.*“ sind Ortsangaben; „*IBM*“ und „*Microsoft*“ sind Organisationen; auf „*Mr.*“ Folgt üblicherweise ein Personenname. Mit Hilfe eines Syntax-Parsers werden zwei spezifische Konstruktionen aus einem nicht-annotierten Korpus extrahiert: Zum einen als Eigennamen getaggte Nomen in Singular-NPs, die durch eine weitere NP postmodifiziert werden („*…says Mr. Cooper, a vice president of…*“) und Eigennamen in PPs, die an einer NP angebunden sind („*…on a federally funded sewage plant in Georgia…*“). Nur die Konstruktionen dieser Art werden als Lerninstanzen für unüberwachtes Lernen repräsentiert, wobei die externe Evidenz, das Nomen im Kontext, getrennt von der internen Evidenz, den Oberflächenmerkmalen des Wortes, abgelegt wird. Die besten Ergebnisse werden mit einem Verfahren erzielt, dass stark an [RILOFF & JONES 1999] erinnert. Die Instanzen werden aufgrund des initialen Wissens klassifiziert. Anhand dieser Daten werden die Merkmale der internen Evidenz gewichtet, um damit erneut alle Instanzen zu klassifizieren. An ihnen werden die Kontextmerkmale

gewichtet, um damit wieder alle Instanzen zu klassifizieren. Das Verfahren wird iterativ benutzt, wobei immer nur die besten Merkmale berücksichtigt werden. Auf denselben Instanzen wird auch der Einsatz eines Boosting-Verfahrens präsentiert, wobei eine modifizierte Version von AdaBoost ([FREUND & SCHAPIRE 1997]) eingesetzt wird. Die Idee dabei ist, zwei Klassifizierer zu benutzen, wobei einer auf interner, der andere auf externer Evidenz basiert. Die unterschiedliche Sicht auf die Instanzen wird eingesetzt, um die Klassifizierer wechselseitig zu optimieren. Das Vorgehen basiert auf der Annahme, dass die Zielsetzung der gegenseitigen Übereinstimmung auf möglichst vielen Instanzen die Klassifizierer optimiert.

Ein NER-System, welches aufgrund des strikt zeichenorientierten Ansatzes prinzipiell auf alle Sprachen anwendbar ist, wird in [CUCERZAN & YAROWSKY 1999] vorgestellt. In [CUCERZAN & YAROWSKY 2002] wird dieses System für spanische und niederländische Texte eingesetzt. Während die Arbeiten von ([RILOFF & JONES 1999], [THELEN & RILOFF 2002]) und [COLLINS & SINGER 1999] auf der strikten syntaktischen Reihenfolge des Englischen aufsetzen und dazu einen syntaktischen Parser benötigen, schlagen [CUCERZAN & YAROWSKY 1999] einen Ansatz vor, der die Instanzen einzig zeichenbasiert repräsentiert und damit unabhängig von einzelnen Sprachen ist. Zwar erlauben alle in den Experimenten verwendeten Sprachen eine leerzeichenbasierte Wortsegmentierung, doch wird die Möglichkeit der rein zeichenbasierten Repräsentation für asiatische Sprachen betont. Instanzen sind diejenigen Wörter, die möglicherweise NEs sind und diese werden repräsentiert durch den linken und rechten Kontext, und den Beginn und das Ende des Wortes. Diese vier Evidenzen werden durch vier Klassifizierer modelliert, welche durch Bootstrapping mittels Expectation Maximization ([DEMPSTER ET AL. 1977]) optimiert werden. Dazu wird eine Liste mit einigen hundert verlässlichen Instanzen der Klassen benutzt, um alle Textvorkommen dieser Instanzen zu klassifizieren. Darauf lernen die vier Klassifizierer und werden auf alle Instanzen angewendet. Stimmen mehr als zwei Klassifizierer bei einer Instanz überein, so wird diese den verlässlichen Instanzen hinzugefügt und erneut zur Klassifikation verwendet. In [CUCERZAN & YAROWSKY 1999] werden für Experimente mit den Klassen Vornamen, Nachnamen und Ortsangaben für die Sprachen Rumänisch, Englisch, Griechisch, Hindi und Türkisch beschrieben. Für den NE Shared Task mit spanischen und niederländischen Texten ([TJONG KIM SANG 2002b]) wurde das System leicht modifiziert. Anstelle der von Hand vorgefertigten Listen wurden die zur Verfügung gestellten annotierten Daten zum Training benutzt. Außerdem bestanden die zu erkennenden NEs oft aus mehreren Wörtern, so dass zusätzliche

Heuristiken erforderlich waren, um mit inkonsistenten Klassifikationen (z.B. eine Ortsangabe eingebettet in Organisationen) umzugehen. Der Ansatz erzielte im Vergleich zu den anderen Beiträgen des Shared Tasks gute Ergebnisse.

Ein Verfahren zum Sammeln von Namen bzw. Namensbestandteilen aus nicht-annotierten deutschen Texten wird für Vor- und Nachnamen in [QUASTHOFF & BIEMANN 2002], und für Firmennamen und Rechtskürzel in [BIEMANN ET AL. 2003]) beschrieben. Benutzt werden kleine Listen mit bekannten Vornamen, Nachnamen, Namensbegrenzern, Titeln und Berufsbezeichnungen für Personen, und Rechtsformenkürzeln für Firmen. Die Listen werden eingesetzt, um mit einem kleinen Set handgeschriebener Regeln neue Namen zu lernen. Die gefundenen Namen bzw. ihre Bestandteile sind Kandidaten zur Aufnahme in die verschiedenen Listen, müssen jedoch vorher verifiziert werden. Die Verifikation basiert grundsätzlich auf einem Schwellenwert, der festlegt, wie oft ein Bestandteil als Kandidat gefunden wurde, im Verhältnis zur gesamten Frequenz des Wortes. Um bei Firmennamen die Extraktion von nicht dazu gehörigen vorangehenden Nomen zu verhindern, werden zusätzliche Frequenzschwellen und Heuristiken angewendet. Um Vornamen von Titel und Berufsbezeichnung zu unterscheiden, wird ein überwachter Entscheidungsbaum eingesetzt, der die beiden Klassen anhand von Substrings unterscheidet. [QUASTHOFF & BIEMANN 2002] zeigen auch, wie der Ansatz benutzt werden kann, um unüberwacht Regeln zur Personennamenerkennung zu lernen.

[LIN ET AL. 2003] präsentieren mehrere Experimente mit einem Ansatz, der in vielen Punkten [RILOFF & JONES 1999] nachempfunden ist. Mit einer kleinen Liste von Namen der gesuchten Klasse wird ein nicht annotiertes Korpus klassifiziert. Bei jedem gefundenen Vorkommen wird der linke und der rechte Kontext von jeweils drei Wörtern extrahiert. In den extrahierten Kontexten werden einige Wörter durch Wildcards ersetzt und die Kontexte werden anhand der positiven Treffer, der unbekannten und der negativen Treffer gewertet. Als negative Treffer gelten Fundstellen, die bereits vom Klassifizierer einer anderen Klasse besetzt sind. Nur die am besten gewichteten Kontexte werden akzeptiert und erneut auf das Korpus angewendet, um neue Namen der gesuchten Klasse zu finden. Da meist nur ein linker oder ein rechter Kontext die Grenzen des Namens anzeigen, wird ein POS-Tag basierter NP-Erkenner eingesetzt, um die jeweils fehlende Grenze zu bestimmen. Die gefundenen Namen werden gewichtet und die besten Einträge werden benutzt, um iterativ nach neuen Kontexten und Namen zu suchen. [LIN ET AL. 2003] betonen insbesondere die Bedeutung von negativen Beispielen für eine Klasse. Als negative Beispiele dienen dabei die Namen anderer, simultan

gelernter Klassen. Beim Lernen von Krankheiten und Ortsangaben in einem Korpus mit Texten über den Ausbruch infektiöser Krankheiten wird gezeigt, dass das simultane Lernen der beiden Klassen bessere Ergebnisse als das isolierte Lernen von Krankheiten liefert. Noch besser werden die Ergebnisse, wenn zusätzliche Klassen wie Personen, Datumsangaben oder Namen von Reports gelernt werden, um zusätzliche, negative Beispiele zu gewinnen. In weiteren Experimenten werden die klassischen MUC-Kategorien in englischen und in chinesischen Texten gelernt, wobei die Aufgabe bei chinesischen Namen auf Einwortnamen begrenzt ist und mit Lexika unterstützt wird.

Die Evaluation des Einsatzes nicht-annotierter Daten

Die Evaluationen des Einsatzes nicht-annotierter Daten sind aufgrund unterschiedlicher Evaluationsmethoden und Korpora kaum vergleichbar. Die Instanzen-basierte Evaluation entspricht der klassischen NER-Evaluation, bei der anhand von Testdaten Precision und Recall gemessen werden. Im Gegensatz dazu bewertet die Typ-basierte Evaluation die Anzahl und Güte der NEs, die aus einem nicht-annotierten Korpus gewonnen werden.

Bei der Instanzen-basierten Evaluation kann das Verfahren direkt auf einen Text angewandt oder mit einem herkömmlichen ML- oder regelbasierten Verfahren kombiniert werden. Dazu werden die gelernten NEs und die gelernten Kontexte als Ressource zur Unterstützung und Erweiterung des eigentlichen Systems eingesetzt. Die Instanzen-basierte Evaluation des Einsatzes nicht-annotierter Daten ist in [BIEMANN ET AL. 2003] und in [LIN ET AL. 2003] beschrieben.

[BIEMANN ET AL. 2003] messen in einem nicht beschriebenen Korpus eine Precision von 97,5% und einen Recall von 78,2% für Personennamen. Diese Resultate sind beeindruckend und liegen mit einem F-Measure von 86.8 sehr nahe an dem besten regelbasierten System von [VOLK & CLEMATIDE 2001], welches ein F-Measure von 88.9 erreichte. Eine andere Arbeit derselben Autoren ([QUASTHOFF & BIEMANN 2002]) lässt jedoch vermuten, dass das nicht weiter beschriebene Evaluationsmaß kein exaktes Matching-Schema benutzt, sondern auf einer Wort-basierten Bewertung beruht. [LIN ET AL. 2003] evaluieren ihr Verfahren auch auf den MUC-7 Texten und erzielen ein F-Measure von 75.1 für PERSON, 73.7 für ORGANISATION und 84.1 für LOCATION. Damit liegen die Ergebnisse deutlich unter den besten MUC-Beiträgen, deren F-Measure für alle Kategorien höher als 90 ist. Dennoch sind die Ergebnisse sehr interessant, da der Aufwand zur Entwicklung eines solchen Systems äußerst gering ist.

Bei der Typ-basierten Evaluation wird bewertet, wie viele unterschiedliche NEs gefunden werden. Allerdings gestaltet sich sowohl die Messung von Recall als auch von Precision schwierig. Der Recall setzt die Menge der tatsächlich zu findenden NEs voraus, die in einem nicht-annotierten Korpus nicht bekannt ist. [COLLINS & SINGER 1999] zählen den Recall einzig in den vorgegebenen syntaktischen Mustern, was nur eine Tendenz zu Tage fördert, aber keine vergleichbaren Zahlen. [LIN ET AL. 2003] gehen hierzu einen interessanten Weg, indem sie über Lexika und von Hand zusammengestellte Listen einen Ausschnitt eines Goldstandards nachbilden. Der Goldstandard entspricht dabei der Ausgabe eines idealen Systems und anhand eines Ausschnittes davon kann der Recall abgeschätzt und sogar zwischen unterschiedlichen Systemen verglichen werden. Allerdings muss bei der Gestaltung eines Goldstandards auch die Frage der Kurzformen von Namen berücksichtigt werden. Bei der Evaluation der Precision muss mit einer gewissen Unschärfe gerechnet werden, da die gefundenen NEs ohne Kontext evaluiert werden und dies auch für Menschen eine äußerst schwierige Aufgabe ist. Während einige Einträge auch ohne Kontext eindeutig als falsch klassifiziert werden können, sind die Klassifikation von Akronymen, die Beurteilung der Vollständigkeit eines Namens und mögliche Mehrdeutigkeiten ohne Kontext schwierig zu beurteilen.

4.4. Diskussion der Korpus-Adaptivität von NER-Systemen

Der Anpassungsfähigkeit von NER-Systemen an eine neue Aufgabe wird in der einschlägigen Literatur wenig Beachtung geschenkt. Diese für NER-Systeme zentrale Eigenschaft wird im Folgenden durch den Begriff der Korpus-Adaptivität diskutiert. Unter Korpus-Adaptivität wird dabei die Fähigkeit eines Systems zur effizienten Anpassung an ein neues Korpus verstanden. Der Begriff der Adaptivität impliziert oftmals eine autonome, d.h. selbständige Anpassung an eine neue Umgebung. Allerdings erfordert die Anpassung eines NER-Verfahrens an ein neues Korpus meist auch die manuelle Anpassung von Ressourcen. Ein System, welches optimal dazu geeignet ist, diesen Prozess der Anpassung effizient zu gestalten, wird im Folgenden *Korpus-adaptiv* genannt. Im Gegensatz zum oftmals verwendeten Begriff der Adaptivität bezüglich einer Domäne, ist die Adaptivität bezüglich eines Korpus besser geeignet, die Anpassungsfähigkeit eines NER-Systems zu charakterisieren. Zum einen wird der Begriff der Domäne für viele Bereiche angewandt und zeichnet sich deshalb durch eine große Unschärfe aus. Selbst wenn der Begriff Domäne als thematische Domäne genauer spezifiziert wird, beschreibt er die Anforderungen an die

Anpassungsfähigkeit eines NER-Systems nur unzureichend. Schließlich sind in vielen Anwendungen nicht Texte einer, sondern solche mehrerer thematischer Domänen zu verarbeiten. Ein Firmenintranet etwa, welches mittels NER semantisch zu erschließen ist, kann sowohl Dokumente aus der technischen als auch aus der wirtschaftlichen Domäne enthalten und darüber hinaus aus zum Teil informell gehaltenen Emails zu diversen Themen bestehen. Selbst ein homogen erscheinendes Korpus einer Tageszeitung besteht aus den thematischen Domänen Politik, Wirtschaft, Sport etc. Aus Sicht der NER ist deshalb davon auszugehen, dass jede neue Anwendung die Verarbeitung eines anderen Korpus erfordert und die Anpassungsfähigkeit eines Systems durch seine Korpus-Adaptivität zu charakterisieren ist.

Ausgehend von der Korpus-Adaptivität existierender NER-Verfahren werden in diesem Kapitel Eigenschaften diskutiert, welche der Anpassung auf ein neues Korpus dienlich sind bzw. diese behindern. Daraus können die Anforderungen an ein Korpus-adaptives System abgeleitet werden. Diese Anforderungen bilden die Grundlage zur Entwicklung des Korpus-adaptiven Systems, welches in Kapitel 5. beschrieben und in Kapitel 6. evaluiert wird.

4.4.1. Kriterien der Korpus-Adaptivität

Es gibt bisher keine systematischen Untersuchungen zur Adaptivität existierender NER-Verfahren. Allerdings kann angenommen werden, dass alle Systeme auf dem Korpus, auf dem sie entwickelt und evaluiert wurden, sehr viel bessere Resultate erzielen als bei der Anwendung auf andere Texte. Gestützt wird dies unter anderem durch die von [MANDL ET AL. 2005] durchgeführte Evaluation von NER-Systemen anhand der Topics des Cross Language Evaluation Forum (CLEF). CLEF bietet eine Infrastruktur zur Bewertung von Information Retrieval Systemen in mehrsprachigen Kontexten. Eine Benutzeranfrage wird als Topic bezeichnet und die mehrsprachige Sammlung dieser Topics wurde mit NEs der Kategorien Person, Ortsangabe und Organisation ausgezeichnet. Sowohl die zwei regelbasierten Systeme GATE (General Archtecture for Text Engineering, [CUNNINGHAM 2000], [CUNNINGHAM ET AL. 2003]) und KIM ([ONTOTEXT 2003]) als auch das lernbasierte System LingPipe[1] sind mit NER-Modellen für mehrere Sprachen verfügbar und wurden zur Annotation der Topics

[1] Das System LingPipe ermöglicht neben weiteren Verfahren zur Analyse natürlichsprachlicher Eingaben auch die Identifizierung von NEs. Weitere Beschreibungen des Systems sind nicht verfügbar. Das System ist unter http://www.alias-i.com/lingpipe/index.html [14.7.2005] erhältlich.

angewandt. Die Systeme sind vorrangig für englische Texte optimiert, so dass die darauf erzielten Ergebnisse am aussagekräftigsten sind. Die besten Erkennungsraten bei Personen (F-Maß 0,53) und Organisationen (F-Maß 0,63) liegen weit unter den MUC-Ergebnissen (vgl. Kap. 4.1.1). Einzig bei Ortsangaben ist das beste Ergebnis (F-Maß 0,87) nahezu mit den MUC-Ergebnissen vergleichbar. Das lernbasierte System LingPipe war den beiden regelbasierten Systemen durchgehend unterlegen. Dies ist mit großer Sicherheit auf die sehr viel geringere Ausstattung des lernbasierten Systems mit Ressourcen zurückzuführen und keinesfalls als allgemeiner Befund zu bewerten. Eine detaillierte Systembeschreibung von LingPipe liegt zwar nicht vor, doch ist aufgrund der Angaben auf der Website davon auszugehen, dass die frei erhältlichen NER-Modelle ohne weitere Ressourcen auf einem annotierten Korpus trainiert worden sind und als Hinweis darauf dienen, dass eine Optimierung der Modelle als kostenpflichtige Dienstleistung zur Verfügung steht.

Auch die Arbeit von [AI ET AL. 2003] weist auf eine mangelnde Korpus-Adaptivität existierender NER-Systeme hin. Im Rahmen eines Informationsextraktionsprojekts mit deutschen Texten wurde die NER-Komponente des SPPC ([PISKORKSI & NEUMANN 2000], [NEUMANN & PISKORKSI 2002], siehe auch Kap. 4.1.2) evaluiert. Die Erkennungsraten bei Personen und Organisationen lagen sowohl bei Precision als auch bei Recall weit unter den in [NEUMANN & PISKORKSI 2002] gemessenen Werten. Im Bereich Ortsangaben liegt das F-Maß mit 0.63 zwar ebenfalls deutlich unter den angegebenen Werten, aber hierbei handelt es sich um ein reines Recall-Problem, da die Precision mit 96% außerordentlich hoch ist.

Ausgehend von diesen Befunden ist zu diskutieren, welche Bereiche eines NER-Systems einer Anpassung bedürfen, mit welchem Aufwand eine solche verbunden ist und von welchen Faktoren eine effiziente Anpassung abhängt. Um eine vergleichbare Einschätzung der lern- und der regelbasierten Systeme hinsichtlich ihrer Korpus-Adaptivität zu erhalten, wird die folgende Struktur in der Betrachtungsweise der anzupassenden Bereiche gewählt:

- Die Anpassung manuell erstellter Regeln bzw. die Annotation eines Trainingskorpus
- Die Anpassung lexikalischer Ressourcen, also der Namens- und Triggerlisten
- Die Anpassung externer Komponenten, auf die ein NER-System zugreift.

4.4.2. Die Anpassung von regel- und lernbasierten Systemen

Bei der Frage, ob regelbasierte oder lernbasierte Systeme leichter anzupassen sind, ist in der bisherigen Forschung keine einhellige Meinung zu erkennen. Sehr oft wird sowohl die Wahl des Lern- als auch des Regelparadigmas damit begründet, dass das jeweils andere Paradigma

zu stark Korpus-abhängig sei (z.B. [NEUMANN & PISKORKSI 2002] über lernbasierte Systeme) bzw. der Anpassungsaufwand viel zu groß sei (z.B. [BIKEL ET AL. 1999] über regelbasierte Systeme). Die folgende Diskussion beschränkt sich auf die Anpassbarkeit von Regeln und stellt dieser den Aufwand einer manuellen Korpusannotation gegenüber. Weitere Systembereiche, die eine Anpassung erfordern, betreffen sowohl die regel- als auch die lernbasierten Verfahren und werden weiter unten besprochen.

Handgeschriebene Regeln zur Identifizierung von NEs werden für ein bestimmtes Korpus entwickelt bzw. auf einem Testkorpus evaluiert. Genügt die Erkennungsleistung auf dem Testkorpus, so gilt der Prozess des Regelschreibens als abgeschlossen. Dadurch ist das Spektrum der abgedeckten sprachlichen Phänomene gewissermaßen durch die im Testkorpus vorkommenden Phänomene bestimmt. Selbstverständlich wird der ambitionierte Regelschreiber versuchen, möglichst viele Phänomene mit Regeln zu modellieren, doch für eine systematische Modellierung ist die Vielfalt der sprachlicher Einheiten und deren Kombinationsmöglichkeiten einfach zu groß. Deshalb erfordert die Anwendung eines regelbasierten Verfahrens auf ein neues Korpus üblicherweise eine Überarbeitung bzw. Optimierung der Regeln.

Aus denselben Gründen müssen lernbasierte Systeme bei der Anwendung auf ein neues Korpus angepasst werden. Das NER-Modell ist anhand eines Trainingskorpus gelernt worden und wird bei der Anwendung auf ein neues Korpus immer dann zu falschen NE-Annotationen neigen, wenn es mit sprachlichen Phänomenen konfrontiert wird, die im Trainingskorpus nicht vorkamen. Im Gegensatz zu regelbasierten Modellen kann ein lernbasiertes Modell kaum manuell optimiert werden, sondern es bedarf eines neuen oder zumindest eines erweiterten Trainingskorpus, was manuelle Annotation erfordert.

Davon ausgehend, dass die Überarbeitung der Regeln und die Annotation zusätzlicher oder neuer Trainingsdaten dasselbe Ziel hat, ist es angemessen, die beiden Anpassungen zu vergleichen.

Die Anpassung handgeschriebener NE-Regeln

Um den Aufwand zur Anpassung handgeschriebener Regeln abzuschätzen, ist es hilfreich, den Umfang solcher Regelwerke und den Zeitaufwand zur Erstellung zu kennen. Allerdings legen nur wenige Arbeiten genauere Informationen hierüber offen.

Anhand der Beschreibungen zweier MUC-7 Systeme (vgl. Abschnitt 4.1.1) lassen sich zumindest Größenordnungen ablesen. Das System von [HUMPHREYS ET AL. 1998], welches

die Grundlage zur Entwicklung von GATE ([CUNNINGHAM 2000], [CUNNINGHAM ET AL. 2003]) war, enthält eine NE-Grammatik mit etwa 150 Regeln. Weniger Regeln, nämlich etwas mehr als 100, enthält das FACILE-System ([BLACK ET AL. 1998]), es erzielt jedoch auch verhältnismäßig schwache Ergebnisse. Eine interessante Bemerkung findet sich dazu bei [BORTHWICK 1999]. Er stellt die Frage, warum das FACILE-System 10 Punkte weniger erzielte als das zweitplazierte System von Isoquest ([KRUPKA & HAUSMANN 1998]). Beide Systeme haben einen vergleichbaren, regelbasierten Aufbau und nutzen ähnliche Ressourcen. Darüber hinaus hätten beide einen Zeitaufwand von einem Monat zur Systementwicklung angegeben. [BORTHWICK 1999] geht davon aus, dass der kommerzielle Einsatz des Systems von Isoquest und die damit verbundenen Überarbeitungen in der Zeit zwischen MUC-6 und MUC-7 zu diesem großen Unterschied geführt hätten. Dies erlaubt den Entwicklern zum einen auf einen großen Erfahrungsschatz in der Systemanpassung, zum anderen auf einen breiten Fundus an Regeln zurückzugreifen. Um die Geschwindigkeit ihres Systems zu evaluieren, testeten [KRUPKA & HAUSMANN 1998] ihr System auch mit nur 20% ihrer Regeln und kommen damit auf Werte, die dem FACILE-System vergleichbar sind. Anhand dieser Angaben lässt sich Folgendes ableiten:

- Der Aufwand zur Entwicklung eines regelbasierten Systems beträgt mehr als einen Monat.
- Die Anzahl der Regeln für ein performantes NER-System übersteigt 100 Regeln ganz deutlich. Ohne eine simplifizierende Korrelation zwischen der Regelanzahl und der Systemleistung behaupten zu wollen, ist darauf hinzuweisen, dass das System von [HUMPHREYS ET AL. 1998] mit 150 Regeln Leistungen zwischen FACILE und dem von Isoquest erzielt hat.
- Wie das Beispiel von Isoquest zeigt, können existierende Regeln für neue NER-Aufgaben wiederverwendet werden.

[PASTRA ET AL. 2002] behandeln in ihrer Arbeit explizit die Frage der Wiederverwendung von NE-Grammatiken im Umfeld von GATE, allerdings mit dem Fokus auf unterschiedliche Repräsentationsformalismen. Um die im Rahmen des CONCERTO-Projekts [BERTINO ET AL. 1999] entwickelte NE-Grammatik für GATE nutzbar zu machen, wird der Prozess der Übertragung in den GATE-Regelformalismus beschrieben. Die NE-Regeln für die CONCERTO-Grammatik sind innerhalb von zwei Monaten durch Anpassung der FACILE-Grammatik entstanden. Die Übertragung in den GATE-Formalismus konnte innerhalb einer Woche durchgeführt werden, was an sich eine große Zeitersparnis bedeutet, aber nur die

Übertragung und nicht die Anpassung an ein neues Korpus umfasst. Überhaupt stellt sich die Frage, warum die existierenden Regeln des GATE-Vorgängers ([HUMPHREYS ET AL. 1998]) durch die Regeln von CONCERTO ersetzt und nicht etwa ergänzt wurden, obwohl beide für dieselben Kategorien in englischen Texten entwickelt wurden. Der Schluss liegt nahe, dass NE-Regeln zwar durchaus wieder verwendbar sind, aber nur, wenn die Person des Regelschreibers auch für die Wiederverwendung bzw. Anpassung zur Verfügung steht. Ist diese Person nicht mehr Teil des Teams, so ist es außerordentlich schwierig, sich in das existierende Regelwerk einzuarbeiten und die vielen Abhängigkeiten zwischen den Regeln zu kennen und zu berücksichtigen.

Ein wichtiger Aspekt zur Abschätzung des Aufwandes ist die Abhängigkeit der Anzahl der zu erkennenden Klassen zum Zeitaufwand der Regelerstellung. Für regelbasierte Systeme ist dieser Zusammenhang grundsätzlich linear. Allerdings gilt dies nur, wenn die Erkennung der Klassen einen vergleichbaren Schwierigkeitsgrad aufweist. Wie [PALMER & DAY 1997] an den MUC Kategorien Zeit- und Datumsangaben und Mengenangaben gezeigt haben, gibt es NE-Klassen, für welche die Regelerstellung mit äußerst geringem Aufwand verbunden ist.

Die Annotation von Trainingsdaten

Über den Aufwand der manuellen Erstellung eines annotierten Korpus zum Training eines lernbasierten Systems findet sich in der Literatur kaum Information. Wichtig ist die Frage nach dem Umfang der erforderlichen Trainingsmenge und dem damit verbundenen Aufwand.

Grundsätzlich weisen alle Experimente daraufhin, dass ein mehr an Trainingsdaten auch zu besseren Ergebnissen führt. Allerdings besteht keine lineare Korrelation zwischen den Ergebnissen und der Anzahl der Trainingsinstanzen. [BANKO & BRILL 2001] untersuchten den Einfluss der Anzahl der Trainingsinstanzen auf die Klassifikationsleistung mehrerer Lernverfahren. Aufgrund der Wahl künstlicher Klassifikationsaufgaben war es möglich, klassifiziertes Trainingsmaterial in der bisher unbekannten Größenordnung von einer Milliarde Instanzen zu erzeugen. Eine künstliche Aufgabe ist etwa der Entscheid, ob an einer bestimmten Textstelle die phonetisch ähnlichen „to“, „two“ oder „too“ stehen. Dazu werden aus einem beliebig großes Korpus Kontexte von „to“, „two“ oder „too“ herausgefiltert. Aus den Kontexten werden Instanzen erzeugt, für die ein Lernverfahren voraussagen soll, welches der drei phonetisch ähnlichen Wörter im Kontext vorkam.

Die Ergebnisse legen den Schluss nahe, dass erst eine Verzehnfachung der Trainingsinstanzen zu einem linearen Anstieg der Erkennungsrate führt. In einem solchen Umfang sind

Experimente zur NER nicht möglich; die größten verfügbaren NE-Korpora bleiben immer unter einer Million annotierter Wörter. Dennoch weisen Experimente zur Korrelation von Trainingsdaten und Erkennungsraten auf eine vergleichbare Abhängigkeit hin (beispielsweise [BORTHWICK 1999] oder [ZHOU & SU 2002]). Der Umfang eines mit NEs annotierten Korpus beruht deshalb immer auf einem Kompromiss zwischen vertretbarem Aufwand und der Hoffnung auf bessere Resultate durch mehr Trainingsinstanzen.
Abbildung 4.2 zeigt einige im Rahmen von Evaluationskampagnen annotierte Korpora und deren Umfang. Das größte verfügbare Korpus enthält 500.000 Wörter und ist mit biomedizinischen NEs annotiert.

Evaluationskampagne	Trainingskorpus	Sprache/Domäne
MUC-7 ([MUC-7 1998])	*320.000 Wörter*	*Mit MUC-Kategorien annotierte englische Zeitungstexte*
CoNLL-02 und CoNLL-03 Shared Task ([TJONG KIM SANG 2002b], [TJONG KIM SANG & DE MEULDER 2003])	*220.000 Wörter jeweils*	*Mit MUC-Kategorien annotierte englische, holländische, spanische und deutsche Zeitungstexte*
JNLPBA-2004 Shared Task ([JNLPBA-2004])	*500.000 Wörter*	*Mit biomedizinischen NE-Kategorien annotierte englische Abstracts von Fachartikeln aus der Medline Datenbank ([MEDLINE])*

Abbildung 4.2: Beispiele annotierter Korpora

Zum Annotationsaufwand kursieren in mündlichen Mitteilungen auf Konferenzen unterschiedliche Angaben. In eigenen Experimenten mit studentischen Hilfskräften zeigte sich, dass pro Stunde zwischen 5.000 und 10.000 Wörter annotiert werden können. Das ergäbe für das größte Korpus in Abbildung 4.2 maximal 100 Stunden, also weniger als drei Wochen reine Annotationsarbeit. Zu den 100 Stunden ist ein zusätzlicher Zeitaufwand für die Einarbeitung, die Diskussion unklarer Annotationen (vgl. dazu Kap. 2.) und daraus resultierender Überarbeitungen erforderlich.
Weitgehend unerforscht ist der Einfluss der Annotationsqualität auf die Erkennungsrate. Allerdings kann davon ausgegangen werden, dass eine höhere Konsistenz der Annotationen eine bessere Lerngrundlage bietet. Menschliche Annotationen sind keineswegs perfekt: Zum

einen enthalten sie Fehler, die auf Unaufmerksamkeit beruhen, zum anderen kommt es zu fehlerhaften oder diskussionwürdigen Annotationen durch die unterschiedliche Interpretation einer Textpassage oder der Annotationsrichtlinien. Da im letzten Fall nicht immer von richtig und falsch gesprochen werden kann, wird die Qualität einer menschlichen Annotation anhand der Übereinstimmung einer Annotation mit einer oder mehreren anderen Annotationen desselben Textes gemessen. Diese Übereinstimmung, das so genannte Interannotator-Agreement, betrug bei den MUC-7 Annotationen 97% ([MARSH & PERZANOWSKI 1998]). Bei der Annotation biomedizinischer Entities wurde eine deutlich geringere Übereinstimmung zwischen den Annotatoren festgestellt. In [HIRSCHMAN 2003] wird für die Übereinstimmung ein F-Measure von 0.87, bei [DEMETRIOU & GAIZAUSKAS 2003] ein F-Measure von 0.89 gemessen.

Das Interannotator-Agreement kann nicht nur zur Messung, sondern auch zur Optimierung der Annotationsqualität eingesetzt werden, indem alle unterschiedlich annotierten Stellen einer erneuten Prüfung unterzogen werden. Allerdings verdoppelt sich dadurch der Annotationsaufwand und erfordert einen zweiten Annotator. Werden weitere 40 Stunden zur Entwicklung und Überarbeitung der Annotationsrichtlinien und zum Abgleich sich widersprechender Annotationen angesetzt, so ergibt sich für die Erstellung eines Korpus im Umfang der MUC-7 Trainingsdaten eine Gesamtzeit von etwa einem Personenmonat.

Was den Zusammenhang zwischen Annotationsaufwand und Anzahl der zu erkennenden Klassen betrifft, so ist der Mehraufwand für die Annotation einer zusätzlichen Klasse als gering einzuschätzen. Allerdings trifft dies nur zu, wenn die Anzahl der zu annotierenden Klassen zu Annotationsbeginn feststeht. Eine nachträgliche Annotation einer zusätzlichen Klasse entspricht mehr oder weniger dem Aufwand der erstmaligen Annotation. Darüber hinaus ist für den lernbasierten Ansatz eine gewisse Frequenz der NE-Klasse im Korpus erforderlich. Sehr selten vorkommende NEs eignen sich deshalb schlecht für den lernbasierten Ansatz. Um zu vermeiden, dass der Umfang des zu annotierenden Materials bei selten vorkommenden Klassen ins Unermessliche steigt, kann der Einsatz so genannter Sampling-Verfahren versucht werden. Dazu werden beispielsweise nur ausgewählte Ausschnitte eines Korpus annotiert, oder es werden ausgehend von den wenigen vorkommenden Instanzen künstliche Instanzen erzeugt. Allerdings gibt es im Bereich Sampling für die NER bisher keine bewährten Methoden.

Gegenüberstellung des Zeitaufwandes

Für die Annotation eines Korpus im Umfang des MUC-7 Korpus, das eher zu den größeren Korpora gehört, wurde der erforderliche Zeitaufwand mit etwa einem Monat abgeschätzt. Anhand der Angaben aus diversen MUC-7 Systembeschreibungen kann die Entwicklungszeit einer NE-Grammatik mit mindestens einem Monat angegeben werden. In punkto Zeitaufwand zur Neuerstellung eines Systems ist der lernbasierte dem regelbasierten Ansatz also in etwa gleichzustellen. Dem Argument, dass für manche Aufgaben wie die Erkennung biomedizinischer NEs annähernd doppelt so große Korpora erforderlich seien und sich damit auch der Zeitaufwand zur Annotation verdopple, muss Folgendes entgegengehalten werden: Bisher ist noch kein regelbasiertes System erfolgreich auf das JNLPBA-2004 Shared Task Korpus angewandt worden. Dies liegt sicherlich daran, dass die Entwicklung von Regeln für diese offensichtlich schwierigere Aufgabe ebenfalls sehr viel mehr Zeit benötigt, als der oben abgeschätzte Zeitaufwand von mindestens einem Monat.

Was den Zeitaufwand in Abhängigkeit der Anzahl der zu annotierenden NE-Klassen betrifft, ist ebenfalls keinem der beiden Verfahren ein grundsätzlicher Vorteil zuzusprechen. Wenn es sich um mehrere Klassen mit ähnlicher Korpusfrequenz und ähnlichem Schwierigkeitsgrad der Erkennung handelt, ist der lernbasierte Ansatz attraktiver, da sich der Mehraufwand während der Annotation kaum bemerkbar macht. Der Zeitaufwand zur Regelerstellung verhält sich in diesem Fall linear zur Anzahl der Klassen. Sind einige der Klassen allerdings mit sehr geringer Frequenz im Korpus vertreten, kann der Lernerfolg mangels Instanzen gefährdet sein, so dass dem regelbasierten Ansatz der Vorzug einzuräumen ist. Ebenfalls zu bevorzugen sind regelbasierte Ansätze, wenn eine NE-Klasse aufgrund einer sehr einfachen und leicht zu modellierenden Syntax effizient durch Regeln beschrieben werden können. Für den Einsatz der MUC Kategorien Zeit-, Datums- und Mengenangaben etwa, steht der erforderliche Zeitaufwand zur Annotation in keinem Verhältnis zur Effizienz der Entwicklung von Regeln.

Gegenüberstellung der Wiederverwendbarkeit annotierter Ressourcen und Regeln

Die Wiederverwendung annotierter Korpora zur Anpassung eines Systems an eine neue Aufgabe zeichnet sich durch große Schwierigkeiten aus. Eine leichte Überarbeitung nur der Definition der NE-Klassen, aber auch stilistische, begriffliche und thematische Unterschiede zwischen dem neuen und dem alten Korpus können dazu führen, dass das existierende annotierte Korpus eine unüberschaubare Anzahl von Fehlklassifikationen enthält. Je ähnlicher sich zwei NER-Aufgaben sind, desto größer ist die Chance, das alte NER-Modell direkt zu

verwenden oder nur durch geringfügige Änderungen oder zusätzliche Annotationen im kleineren Umfang anzupassen. Allerdings existieren keine bewährten Verfahren für diesen Vorgang, so dass sich ein langer Trial-and-Error Prozess möglicherweise aufwändiger gestaltet als die vollständige Annotation eines neuen Korpus. Es bietet sich jedoch an, existierende Erkennerkomponenten zur Unterstützung der manuellen Annotation einzusetzen. Eine semi-automatische Annotationsumgebung, die dem Annotator Vorschläge macht, gestaltet den Annotationsprozess effizienter und ermöglicht gleichzeitig eine Evaluation der existierenden Komponenten.

Die Wiederverwendbarkeit von NE-Grammatiken auf einem neuen Korpus ist grundsätzlich möglich. Allerdings erfordert die Wiederverwendung in Verbindung mit der Überarbeitung und Erweiterung die personelle Kontinuität des Regelentwicklerteams. Die Handhabung eines Regelapparates, welche nicht die Erweiterung isolierter Regeln, sondern den Überblick über ein komplexes Wechselspiel aufeinander aufbauender und manchmal konkurrierender Regeln erfordert, ist ohne profunde Kenntnisse der Grammatik unmöglich. Die Einarbeitung ist äußerst zeitaufwendig und wird selten nur durch Dokumentationen erleichtert, da die verständliche Dokumentation eines Regelwerkes an sich einen großen Aufwand darstellt. Steht also die Person des Regelentwicklers nicht mehr für Überarbeitungen zur Verfügung, können die Komponenten zwar weiterverwendet, aber kaum mehr gepflegt werden. Eine mehr oder weniger vollständige Neuerstellung einer NE-Grammatik muss deshalb bei einer notwendigen Überarbeitung in Betracht gezogen werden.

Gegenüberstellung der erforderlichen Qualifikationen

Was die Qualifikationen der Systementwickler betrifft, so ist die manuelle Korpusannotation gegenüber dem Regelschreiben deutlich genügsamer bzw. optimaler geeignet, die erforderliche Expertise in das Modell mit einfließen zu lassen. Die Entwicklung von Regeln kann nur von einem erfahrenen Computerlinguisten geleistet werden, der im Falle fachwissenschaftlicher Texte, zusätzlich über genügend fachwissenschaftliche Kompetenz verfügen muss. Wie es am Beispiel der Erkennung biomedizinischer NEs leicht vorstellbar ist, kann es schwierig und teuer sein, Entwickler mit einer solchen Kombination von Qualifikationen zu finden. Im Gegensatz dazu kann die manuelle Annotation nach einer kurzen Einweisung von Fachexperten oder – im Falle nicht-fachsprachlicher Texte - beispielsweise von Studenten durchgeführt werden. Da der Erwerb des NER-Modells automatisch geschieht, kann auch das Training und die Evaluation des NER-Systems von den

Annotatoren übernommen werden, vorausgesetzt, dass bewährte lernbasierte NER-Verfahren und leicht zu bedienende Werkzeuge zur Verfügung stehen.

4.4.3. Die Anpassung lexikalischer Ressourcen

Wie in Kapitel 3.2.4 eingeführt, umfasst der Begriff der lexikalischen Ressourcen für unsere Zwecke alle möglichen Wortlisten, auf die sowohl lern- als auch regelbasierte NER-Systeme zugreifen. Dazu gehören beispielsweise Trigger-Listen von Funktions- und Berufsbezeichnungen („*Sprecher*“, „*Professor*“), welche die Erkennung eines nachfolgenden Personennamens unterstützen, Listen von Firmenrechtskürzeln („Ltd“, „GmbH“) als Bestandteile von Organisationsnamen oder Listen mit Ortsnamen („Wanne-Eickel“, „Djibouti“).

Die Aufgabe der Listen besteht darin, in Texten vorkommende Listeneinträge aufgrund der Listenkategorie zu klassifizieren und mehrdeutige oder unbekannte Einträge aufgrund der weiteren Triggerlisten zu disambiguieren und zu klassifizieren. Um mit einem vorrangig listenbasierten System bestmögliche Resultate zu erzielen, müssen diese Listen abdeckend und verlässlich sein. Abdeckend bedeutet, dass möglichst alle im Korpus vorkommenden Einträge enthalten sind. Dies wird zum einen durch die ungeheure Anzahl von NEs, aber auch durch Varianten und Kurzformen bekannter NEs (vgl. Kap. 2.2.2) erschwert. Das Kriterium der Verlässlichkeit beschreibt, dass der Anteil der Listeneinträge, die zu einer falschen Klassifikation von Textstellen führen, möglichst gering sein muss.

Weder für die Erstellung noch für die Evaluation des Listeineinsatzes existiert eine etablierte Methodologie. Listen werden aus diversen Quellen zusammengestellt, semi-automatisch gefiltert und anschließend einzig durch die Anzahl der Listeneinträge dokumentiert. Steigt die Erkennungsleistung eines Systems mit diesen Listen, so gilt der Einsatz als gelungen. Diese Betrachtung greift jedoch deutlich zu kurz und vernachlässigt, dass ein Testkorpus nur eine verhältnismäßig kleine Zahl von NEs enthält und zur Evaluation von Listen, deren Umfang oftmals 100.000 Einträge übersteigt, nur bedingt geeignet ist. Darüber hinaus mangelt es an Verfahren, die eine Aktualisierung dieser Ressourcen ermöglichen. Diese ist erforderlich, da in einer realen Anwendung ständig neue Texte zu verarbeiten sind, in denen immer wieder neue, bisher unbekannte NEs, wie etwa Namen von neu gegründeten Firmen oder neu entdeckten Proteinen auftauchen.

Bereits in Kapitel 4.1.1 wurden die interessanten Befunde von [KRUPKA & HAUSMANN 1998] vorgestellt. Diese testeten den Einfluss der Größe der lexikalischen Ressourcen und stellten

fast keinen Einfluss bei der Reduktion der Liste von 110.000 (F-Measure: 91.6) auf 25.000 (F-Measure: 91.45) und einen geringen bei der Reduktion auf 9.000 Einträge (F-Measure 89.13) fest. Gleichzeitig führte die Erweiterung um 42 (!) domänenspezifische Einträge bei der Kategorie Organisation sowohl bei Precision als auch bei Recall zu einer Verbesserung von 6 Punkten, bei den anderen Kategorien zu einer Verbesserung von bis zu 3 Punkten.

Anhand einer kleinen Auswertung über Ortsangaben des deutschsprachigen Testkorpus der CoNLL-03 kann dieser Befund erklärt werden. Das Testkorpus enthält rund 50.000 Wörter und hat damit eine repräsentative Größe, da größere Testkorpora aufgrund des Annotationsaufwandes kaum zum Einsatz kommen. Das Korpus enthält insgesamt 1182 NEs vom Typ Ortsangabe. Diese wiederum bestehen aus 681 Types, d.h. im Durchschnitt kommt jedes Type weniger als zweimal im Korpus vor. Werden nur die zehn bzw. zwanzig häufigsten Types betrachtet, so machen sie aber jeweils 12% bzw. 18% der NE-Vorkommen aus. Das bedeutet zum einen, dass von den mehr als 100.000 Einträgen einer größeren Liste von Ortsnamen nur 681 Einträge überhaupt zum Einsatz kommen. Würde die Liste radikal auf diese 681 Einträge gekürzt, so änderte sich grundsätzlich nichts bei der Anwendung des Systems auf dasselbe Korpus. Einzig der potentielle negative Einfluss irreführender Einträge würde gänzlich eliminiert, wodurch die Ergebnisse möglicherweise sogar besser ausfielen. Zum anderen könnte das Entfernen von 10 bzw. 20 Einträgen die Abdeckung der Liste von 100% auf 88% bzw. 82% senken. Dasselbe gilt natürlich für den umgekehrten Weg, d.h. das Erstellen einer Liste, wo einige wenige Einträge die Abdeckung der Liste massiv erhöhen können. Dies erklärt zum einen die Befunde von [KRUPKA & HAUSMANN 1998], wirft aber auch die Frage auf, wie denn der Einsatz von Listen zu evaluieren ist und was das für den Einsatz und die Anpassbarkeit lexikalischer Ressourcen auf ein neues Korpus bedeutet.

Um diese Fragen zu diskutieren, ist eine detailliertere Betrachtung der lexikalischen Ressourcen notwendig. Sinnvoll erscheint eine Unterscheidung der NE-Type-übergreifenden Einträge, welche Evidenz für mehrere unterschiedliche NEs liefern, von NE-Type-spezifischen Einträgen, welche zur Identifizierung eines spezifischen NE-Types eingesetzt werden. Zu den Type-übergreifenden Einträgen sind beispielsweise Firmenrechtskürzel oder Trigger-Listen zu zählen, zu den Type-spezifischen beispielsweise eine Liste von Ortsnamen. Der Unterschied zwischen Type-spezifisch und Type-übergreifend stellt ein Kontinuum dar, welches auf der spezifischen Seite etwa mit Ortsnamen beginnt, welche nur Hinweise auf eine bestimmte NE liefern und beispielsweise über eine Liste häufiger Vornamen auf der Type-übergreifenden Seite mit Trigger-Listen endet.

Je Type-übergreifender ein Eintrag einer lexikalischen Ressource ist, desto weniger Anpassung an ein Korpus ist erforderlich. Eine Liste von Firmenrechtskürzeln ist in den meisten Korpora zur Erkennung von Organisationen nützlich. Eine Liste afrikanischer Dorfnamen hingegen, wäre bei der Annotation einer amerikanischen Regionalzeitung von fragwürdigem Wert. Im besten Fall kommen die Listeneinträge einfach nicht vor, so dass diese faktisch ohne jeden Effekt bleiben. Allerdings ist es auch denkbar, dass einige afrikanische Dorfnamen irreführend sind, d.h. dass sie die gleiche Form haben wie etwa allgemeine Nomen („Liberty“), Personennamen oder Produktenamen und somit am Satzanfang oder bei Eigennamen an jeder Position im Satz fälschlicherweise auf eine Ortsangabe hindeuten.

Von diesen Überlegungen ausgehend, können die Anforderungen an optimale lexikalische Ressourcen genauer spezifiziert werden: Erforderlich sind Listen, die zu jedem Eintrag die Type-Spezifizität und Informationen über die mögliche Mehrdeutigkeit bereitstellen. Der Prozess der Adaption lexikalischer Ressourcen dieser Art bedeutet die Übernahme der bestehenden Ressourcen unter Beibehaltung und ggf. Erweiterung der Type-übergreifenden Einträge, der Überarbeitung und Erweiterung der spezifischen Einträge und die Aktualisierung der Angaben über mehrdeutige Einträge. Letzteres ist erforderlich, da der Grad der Mehrdeutigkeit eines Eintrages eine korpusspezifische Größe und kein absoluter Wert ist. So ist etwa „*Winterthur*“, der Name einer schweizerischen Stadt und gleichzeitig eine Kurzform für „*Winterthurer Versicherung*“. In der Regionalzeitung der Stadt Winterthur wird „*Winterthur*“ fast immer als Städtenamen verwendet, während in einem versicherungsspezifischen Korpus die Verwendung des Städtenamens kaum vorkommen wird. Allerdings sind diese Informationen auf dem herkömmlichen Weg der Listenerstellung nicht zu erhalten. Der Vorteil umfangreicher Listen liegt in der schnellen Erstellung aus diversen, vorliegenden Quellen. Ein manuelles Einarbeiten dieser Informationen ist selbst bei überschaubaren Listen von einigen tausend Einträgen kaum möglich, da eine Abschätzung dieser Informationen bei vielen Einträgen eine Korpusstudie erfordert, welche zu einem nicht mehr vertretbaren Zeitaufwand führt.

Trotz der aufgezeigten Schwierigkeiten ist der Einsatz von Listen hilfreich und geradezu unverzichtbar. Dennoch resultieren daraus Konsequenzen zum Einsatz und zur Anpassung lexikalischer Ressourcen, die in den meisten existierenden Ansätzen weitestgehend vernachlässigt werden:

- Die Evaluation von weitestgehend listenbasierten Ansätzen ist stark von den verwendeten Listen abhängig. Die verhältnismäßig geringe Anzahl von NEs in einem üblichen Testkorpus und deren Verteilung führt dazu, dass das Hinzufügen oder Entfernen einzelner Einträge zu gänzlich anderen Ergebnissen führt. Systeme sind deshalb eigentlich nur vergleichbar, wenn sie mit denselben lexikalischen Ressourcen bestückt sind, oder gänzlich ohne Type-spezifische Listen eingesetzt werden. Letzteres führt zwar meist zu deutlich schlechteren Ergebnissen, ermöglicht jedoch erst eine vergleichende Evaluation der übrigen Komponenten eines Systems. Jedoch ist dies für regelbasierte Systeme kaum möglich, da diese meist auf stark lexikonbasierten Grammatiken beruhen. Für regelbasierte Systeme ist deshalb zumindest eine aufwändige Dokumentation der verwendeten Ressourcen wünschenswert; allerdings ist dies mit nicht geringem Aufwand verbunden. Außerdem ist eine Aufteilung der Test-Daten in Entwicklungs- und finale Testdaten noch von größerer Wichtigkeit als bei lernbasierten Verfahren.
- Um Korpus-adaptive Systeme zu entwickeln, sind Verfahren notwendig, welche die effiziente Erstellung der benötigten lexikalischen Ressourcen ermöglichen. Ansätze hierzu existieren vor allem im Bereich des Lernens aus nicht-annotierten Daten (vgl. dazu Kapitel 4.3). Werden als nicht-annotierte Daten Texte des zu verarbeitenden Korpus gewählt, so kann damit die Spezifizität der lexikalischen Ressourcen erreicht werden. Nicht-annotierte Daten stellen darüber hinaus eine äußerst attraktive Möglichkeit zur Wartung und Aktualisierung von NER-Systemen dar. Ein Ansatz, der vorrangig auf manuell erstellten Listen basiert, hat abgesehen von der menschlichen Nachkontrolle der Annotationen und der anschließenden Einpflege neuer NEs in die lexikalischen Ressourcen keine Möglichkeiten, ein System aktuell zu halten.
- Die Nützlichkeit semi-automatisch oder manuell erstellter Ressourcen soll keinesfalls in Abrede gestellt werden. Beim Einsatz der NER in konkreten Anwendungen sind sicherlich alle verfügbaren und hilfreichen Ressourcen heranzuziehen. Allerdings zeigt sich auch hier, dass größere Listen nicht zwangsläufig zu besseren Ergebnissen führen, oder wie [KORNAI & THOMPSON 2005] es ausdrücken: „Size doesn't matter". Eine größere Beachtung des Einsatzes und der Anpassung lexikalischer Ressourcen dient dazu, solche kontra-intuitiven Befunde besser zu verstehen und kann dadurch bessere NER-Verfahren ermöglichen. Das manuelle Hinzufügen und Entfernen von Listeneinträgen, um bessere Resultate zu erhalten, ist von größter Wichtigkeit, um in realen Anwendungen eine optimale Performanz zu ermöglichen. In forschungsorientierten Arbeiten zur NER ist

dieses Vorgehen jedoch aufgrund mangelnder Nachvollziehbarkeit und fehlender theoretischer Motiviertheit nur mit Vorsicht einzusetzen.

4.4.4. Die Anpassbarkeit der Generalisierung sprachlicher Einheiten

Die Generalisierung über sprachliche Einheiten ist essentiell für alle Systeme der automatischen Sprachverarbeitung. Zwar geht die Generalisierung mit einem Verlust an spezifischer Information einher, doch scheitert die nicht auf Abstraktion beruhende Repräsentation sprachlicher Einheiten an der Vielfalt der Einheiten und ihrer schier unerschöpflichen Kombinationsmöglichkeiten. Erst die Zusammenfassung nach morphologischen, syntaktischen und semantischen Eigenschaften ermöglicht es, in der Sprache Strukturen zu erkennen, die eine effiziente Modellierung erlauben. In Kapitel 3.2 sind die für die NER wichtigsten Generalisierungen vorgestellt worden. Die semantische Generalisierung erfolgt bei der NER über die Erstellung lexikalischer Ressourcen, deren Korpus-Adaptivität bereits im vorhergehenden Abschnitt 4.4.2 diskutiert wurde. Für die Zusammenfassung linguistischer Einheiten nach morphologischen und syntaktischen Kriterien stehen modellorientierte (siehe Abschnitt 3.2.2) und datenorientierte (siehe Abschnitt 3.2.3) Verfahren zur Verfügung. Modellorientierte Verfahren basieren auf linguistisch anerkannten Sprachmodellen und erfordern einen nicht geringen Implementierungsaufwand. Deshalb werden diese Verfahren in NER-Systemen meist in Form externer Tools integriert. Die datenorientierte Generalisierung hingegen beruht einzig auf der Beobachtung der vorliegenden Daten. Die Zeichenketten weisen Eigenschaften auf, welche eine Klassenbildung ermöglichen, die auf einfach zu extrahierenden Merkmalen wie etwa den folgenden beruhen: beginnt mit einem Großbuchstaben, enthält Ziffern, folgt auf eine bestimmte Zeichenkette, enthält einen bestimmten Substring etc. Generalisierungen dieser Art sind durchaus in der Lage, einige linguistische Regularitäten abzubilden, müssen sich jedoch den Vorwurf gefallen lassen, manche Eigenschaften der Sprache in simplifizierender und unangemessener Weise zu modellieren (vgl. dazu beispielsweise die Diskussion der N-Gramm Modellierung syntaktischer Eigenschaften in Abschnitt 3.2.3).

Aus Sicht der NER ist jedoch nicht die Angemessenheit der Modellierung von Bedeutung, sondern einzig der Einfluss der Generalisierung auf die Leistung des Systems. Um darüber hinaus Korpus-adaptive Verfahren zu entwickeln, müssen die Generalisierungsverfahren auch auf unterschiedliche Korpora mit der gleichen Zuverlässigkeit anwendbar sein. Dabei zeigt sich eine deutliche Überlegenheit datenorientierter Verfahren. Werkzeuge zur modellbasierten

Generalisierung wie etwa Chunker, POS-Tagger oder Systeme zur morphologischen Analyse sind fast durchgehend auf Zeitungskorpora entwickelt und getestet worden. Diese zeichnen sich durch syntaktische Wohlgeformtheit, mehrheitlich korrekte Orthographie und eine nicht-fachsprachliche Lexik aus. Werden diese Werkzeuge auf informelle Texte, mit OCR- oder ASR erfasste Dokumente oder auf fachsprachliche Korpora angewandt, wird dies sehr viel häufiger zu Fehlern, also zu falschen Generalisierungen führen. Datenorientierte Verfahren hingegen sind aufgrund ihrer Einfachheit nahezu ohne Anpassungsaufwand auf alle Korpora anwendbar.

5. Ein Korpus-adaptives NER-System

Ausgehend von der Diskussion der Korpus-Adaptivität in 4.4 wird in diesem Kapitel ein Korpus-adaptives System vorgestellt. In Form von Entwurfsrichtlinien werden in 5.1 die Anforderungen an die Bestandteile und Eigenschaften des Systems spezifiziert. Charakteristisch für den Ansatz ist der Einsatz des Lernens mit Support Vektor Maschinen, die durchgehende Datenorientierung der Modellierung sowie der Verzicht auf manuell erstellte lexikalische Ressourcen. Um trotz der damit einhergehenden Ressourcenarmut hohe Erkennungsraten zu erreichen, werden nicht-annotierte Daten eingesetzt, welche meist ohne Kosten und ohne weiteren Aufwand in großen Mengen verfügbar sind.

Der Einsatz der nicht annotierten Daten ist von der Idee angetrieben, dass ein einfacher Klassifizierer anhand der Auswertung seiner eigenen Klassifikationsentscheide verbessert werden kann. Dazu wird ein sog. *Basisklassifizierer*, der aufgrund von manuell annotierten Daten trainiert worden ist, zur Klassifikation eines umfangreichen nicht-annotierten Korpus eingesetzt. Basierend auf den Klassifikationsentscheiden werden Merkmale berechnet, welche interne und externe Evidenz zur NER ableiten. Mithilfe dieser sog. *erweiterten Merkmale* wird der Basisklassifizierer zum *erweiterten Klassifizierer*. Zeigt dieser erweiterte Klassifizierer eine verbesserte Erkennungsrate, so wird er erneut zur Klassifikation des umfangreichen nicht-annotierten Korpus eingesetzt und die erweiterten Merkmale werden erneut berechnet. Dieser Vorgang wird solange wiederholt, bis keine Verbesserung mehr zu erkennen ist.

Die Ressourcen des Basisklassifizierers beschränken sich auf ein Trainingskorpus, dessen Instanzen mit den sog. Basismerkmalen repräsentiert werden. Alle Basismerkmale beruhen auf einer strikt datenorientierten Generalisierung. Abschnitt 5.2 beschreibt die eingesetzten Techniken der datenorientierten Modellierung. Die Verfahren, welche die erweiterten Merkmale anhand der automatisch annotierten Daten berechnen, werden in 5.3 vorgestellt. Für diese Aufgabe sind zwei unterschiedliche Varianten entwickelt worden: System I, welches sich auf den Erwerb interner Evidenz beschränkt ([RÖSSLER 2004b]) und das nachfolgende System II, welches den ersten Ansatz modifiziert und insbesondere auch den Erwerb externer Evidenz unterstützt ([RÖSSLER & MORIK 2005]). Ein Überblick über das Gesamtsystem und die einzelnen Verarbeitungsschritte befindet sich in Abschnitt 5.4.

5.1. Entwurfsrichtlinien für ein Korpus-adaptives System

Die Entwicklung eines NER-Systems erfordert vor der Implementation eine Anzahl zentraler Entscheidungen, um den Grundriss des Systems zu definieren. Die Eckpunkte des zu entwickelnden Systems werden im Folgenden festgelegt.

Das Modellierungsparadigma - lernbasiert

Das regelbasierte und das lernbasierte Modellieren benötigen einen vergleichbaren Entwicklungsaufwand. Regelbasierte Systeme sind bei ähnlichen NER-Aufgaben leichter auf ein neues Korpus anzupassen. Allerdings ist dazu hoch qualifiziertes Personal erforderlich, welches bei einem Ausscheiden aus dem Entwicklerteam kaum ersetzt werden kann. Dies kann dazu führen, dass eine existierende NE-Grammatik zwar benutzbar ist, aber nicht mehr gewartet und erweitert werden kann. Das im Vergleich zur Anpassung einer Grammatik aufwändige Annotieren von Daten hat bei lernbasierten Ansätzen darüber hinaus den Vorteil, dass es von Bearbeitern mit geringen linguistischen Qualifikationen geleistet werden kann. Im Falle fachsprachlicher Korpora bietet die manuelle Annotation durch einen Domänenexperten eine effiziente Schnittstelle für das erforderliche Expertenwissen. Bei regelbasierten Systemen hingegen muss der Domänenexperte sein Wissen über die zu annotierenden NEs explizieren, entweder um es einem Regelschreiber zu vermitteln oder um es selbst in Regeln zu erfassen.

Der Einsatz lexikalischer Ressourcen – keine manuell erstellten Ressourcen

Ohne die Nützlichkeit manuell erstellter Ressourcen in Abrede zu stellen, ist es sinnvoll, einen korpus-adaptiven Ansatz so weit als möglich auf automatisch erzeugbare Ressourcen zu beschränken. Die automatische Ableitung von Evidenz mithilfe annotierter und nicht-annotierter Daten ist eine attraktive Möglichkeit, korpus-spezifische Ressourcen zu erstellen. Der Einsatz nicht-annotierter Daten bietet darüber hinaus eine Lösung für die Wartung und Aktualisierung von NER-Systemen. Der Einsatz manuell erstellter Ressourcen beispielsweise im Rahmen einer Annotationsanwendung wird damit keineswegs prinzipiell abgelehnt, wird im Rahmen dieser Arbeit aber unter dem Gesichtspunkt der Adaptivität nicht weiter verfolgt.

Die Generalisierung linguistischer Einheiten – datenorientiert

Die datenorientierte Generalisierung zeichnet sich durch ihre Korpus-Adaptivität bzw. der grundsätzlichen Verfügbarkeit für alle textbasierten NER-Anwendungen aus. Bei Korpora, für deren Texte geeignete Werkzeuge zur Erzeugung linguistisch modellierter

Generalisierungen zur Verfügung stehen, stellt sich sicherlich die Frage, ob diese nicht zu einer besseren Erkennungsrate führen. Da hierüber systematische Evaluationen fehlen, lässt sich dies nur in eigenen Experimenten überprüfen. Analog zum Umgang mit manuell erstellten lexikalischen Ressourcen, wird hier eine ähnliche Strategie verfolgt: Die Entwicklung eines Korpus-adaptiven Systems kann nur durch den Verzicht modellorientierter Generalisierung erreicht werden. Besteht für eine NER-Aufgabe die Möglichkeit, auf linguistische Klassen zurückzugreifen, so spricht nichts gegen deren Einsatz. Um dies zu ermöglichen, muss ein NER-Verfahren allerdings auch in der Lage sein, eine Vielzahl von Evidenzen zu berücksichtigten, was insbesondere die Wahl des Lernverfahrens beeinflusst.

5.2. Basismerkmale und Kontextmodellierung

Die Auswahl der Merkmale ist neben der Wahl des Lernalgorithmus der zentralste Aspekt eines lernbasierten Ansatzes. In diesem Abschnitt werden die sog. Basismerkmale und die jeweils gewählte Konfiguration beschrieben. Die Basismerkmale stehen in Abgrenzung zu den sog. erweiterten Merkmalen, welche erst durch die Auswertung nicht-annotierter Daten gewonnen werden und in Abschnitt 5.3 beschrieben werden. Die Basismerkmale müssen die Eigenschaften eines Wortes, aber auch die aktuelle Verwendung des Wortes abbilden. Die Verwendung eines Wortes kann durch eine geeignete Modellierung des Kontexts des Wortvorkommens beschrieben werden.

In Abschnitt 5.2.1 werden zwei unterschiedliche Verfahren zur Modellierung des Kontexts diskutiert. Das Standardverfahren der N-Gram Modellierung, bei der für jedes Wort ein Kontextfenster der Größe N benutzt wird, wird als *traditionelle N-Gram Modellierung* bezeichnet. Dies steht im Gegensatz zur *erweiterten N-Gram Modellierung*, einem innovativen Ansatz, welcher erstmals in [RÖSSLER & MORIK 2005] beschrieben wurde. Direkt damit verbunden ist auch die Frage, welche Einheiten als Instanz gewählt werden. Abschnitt 5.2.2 beschreibt Merkmale, welche zur Generalisierung der Wortoberfläche eingesetzt werden und Abschnitt 5.2.3 stellt eine Repräsentation von Wörtern mittels Substring-Zerlegung vor, welche sowohl Spezifisches über Wortformen, als auch Ähnlichkeiten zwischen unterschiedlichen Wortformen erfasst. Der letzte Abschnitt bietet einen Überblick über die verwendeten Basismerkmale und exemplarische Abbildungen von Instanzen.

5.2.1. Die Modellierung der Instanz und des Kontexts

Da die NER nicht Wörter, sondern Wortvorkommen klassifiziert, muss eine zu klassifizierende Instanz in ihrem Kontext modelliert werden. Wie in lernbasierten Ansätzen zur NER üblich, wird eine wortweise Klassifikation in Verbindung mit der *traditionellen N-Gram Modellierung* des Kontexts gewählt. Um dem Phänomen Rechnung zu tragen, dass viele NEs aus mehreren Wörtern bestehen, wurden *erweiterte N-Gramme* entwickelt ([RÖSSLER & MORIK 2005]), welche die wortweise Klassifikation durch die Klassifikation von Wortsequenzen ersetzt. Ein Vergleich zwischen diesen alternativen Möglichkeiten der Kontextmodellierung findet sich in Abschnitt 6.1.5.

Aufgrund experimenteller Resultate (vgl. Kap. 6) wurde für die einfache N-Gram Modellierung ein Fenster der Größe sechs gewählt, welches die drei vorangehenden, das zu klassifizierende und die zwei nachfolgenden Wörter umfasst. Die wortweise Klassifikation eines Satzes wird in Abbildung 5.1 exemplifiziert.

Satz: *Wie der Sanitäter Peter Hammer mitteilte, wird das Kind noch heute zurückkehren.*						
Wort-3	**Wort-2**	**Wort-1**	**Fokus**	**Wort+1**	**Wort+2**	**Klasse**
…	…	<satz>	Wie	der	Sanitäter	**O**
<satz>	Wie	der	Sanitäter	Peter	Hammer	**O**
Wie	der	Sanitäter	Peter	Hammer	mitteilte	**I-PER**
der	Sanitäter	Peter	Hammer	mitteilte	,	**I-PER**
,	wird	das	Kind	noch	heute	**O**

Abbildung 5.1: Die Klassifikation mit N-Gram modelliertem Kontext (traditionell)

Wie an Abbildung 5.1 deutlich wird, wird der Satz wortweise durchlaufen, um bei jeder potentiellen NE eine zu klassifizierende Instanz zu erzeugen. Die Festlegung, was eine potentielle NE ist, ist nicht ganz trivial. Grundsätzlich sind erstmal alle großgeschriebenen Einträge, damit auch die Stoppwörter am Satzanfang, dazu zu zählen. Darüber hinaus führt die wortweise Klassifikation auch dazu, dass kleingeschriebene NE-Bestandteile (beispielsweise „Frankfurt *am* Main“) ebenfalls zu klassifizieren sind. Die Frage ist bedeutsam, weil dadurch der Anteil der positiven Instanzen zur Gesamtanzahl der Instanzen festgelegt wird und weil bei einer Beschränkung auf großgeschriebene Einheiten bereits im Vorfeld ein gewisser Klassifizierungsfehler in Kauf genommen wird, der nur durch

Nachbearbeitung behoben werden kann. Darüber hinaus beeinflusst die Anzahl der erzeugten Instanzen die Geschwindigkeit des Verfahrens. Bereits in Vorstudien hat es sich bewährt, das Trainingskorpus nach Einheiten zu filtern, die nicht mit einem Großbuchstaben beginnen, aber dennoch als Teil einer NE vorkommen. Die so gesammelten Einträge bilden zusammen mit allen großgeschriebenen Wörtern die Menge der Token, die als potentielle NEs betrachtet werden und die deshalb als zu klassifizierende Einheiten behandelt werden. Dieses Vorgehen stellt einen Kompromiss dar zwischen der aufwändigen und meist überflüssigen Analyse aller, also auch der nicht-großgeschriebenen Token und der Inkaufnahme einer geringen Fehlerquote.

Die wortweise Verarbeitung ist allerdings nicht die einzige Möglichkeit der datenorientierten Kontextmodellierung. Insbesondere wird dabei der Phrasen-Charakter von NEs vernachlässigt, also das Phänomen, dass viele NEs aus einer Sequenz von Wörtern und nicht aus einem isolierten Wort bestehen. Die gleichzeitige Betrachtung einer Sequenz als potentielle NE und der sie umgebende Kontext kann mit der in [RÖSSLER & MORIK 2005] vorgeschlagenen erweiterten N-Gram Modellierung durchgeführt werden. Dabei werden alle Wortsequenzen, die möglicherweise eine NE darstellen, als Instanzen erzeugt. Abbildung 5.2 zeigt die Erzeugung durch die dynamische Expansion einer Instanz. Das Beispiel beruht auf der vereinfachenden Annahme, dass nur großgeschriebene Einträge eine NE-Sequenz formen können. Deckt eine Sequenz nur einen Teil der NE ab, so muss festgelegt werden, ob dies als positive oder negative Instanz der Klasse betrachtet wird. Um das Lernen von Sequenzen und damit die korrekte Identifikation von Beginn und Ende einer NE zu betonen, wurde festgelegt, dass nur die gesamte NE-Sequenz, aber nicht Teile davon als Instanz der NE-Klasse betrachtet werden.

Satz: *Wie der Sanitäter Peter Hammer mitteilte, wird das Kind noch heute zurückkehren.*						
Wort-3	**Wort-2**	**Wort-1**	**Fokus**	**Wort+1**	**Wort+2**	**Klasse**
…	…	<satz>	Wie	der	Sanitäter	**O**
<satz>	Wie	der	Sanitäter	Peter	Hammer	**O**
<satz>	Wie	der	Sanitäter Peter	Hammer	mitteilte	**O**
<satz>	Wie	der	Sanitäter Peter Hammer	mitteilte	,	**O**
Wie	der	Sanitäter	Peter	Hammer	mitteilte	**O**
Wie	der	Sanitäter	Peter Hammer	mitteilte	,	**I-PER**
der	Sanitäter	Peter	Hammer	mitteilte	,	**O**
,	wird	das	Kind	noch	heute	**O**

Abbildung 5.2: Die erweiterte N-Gram Modellierung eines Beispielsatzes

Da bei der erweiterten N-Gram Modellierung nicht mehr nur ein Wort, sondern eine ganze Wortsequenz im Klassifikationsfokus ist, muss festgelegt werden, wie die Merkmale aller Teile einer Wortsequenz kombiniert werden. So kann eine exemplarische Instanz mit der Sequenz „*Peter Müller*“ im Klassifikationsfokus ganz unterschiedlich kodiert werden. Beispielsweise können jeweils individuelle Merkmale für das erste und das zweite Wort erzeugt werden. Würde die Sequenz „*Johann Peter Müller*“ analog repräsentiert, also mit Merkmalen für das erste, das zweite und das dritte Wort, so erfasste die Repräsentation keinerlei Ähnlichkeiten zwischen den beiden Instanzen. Deswegen wurde eine Repräsentation gewählt, die innerhalb einer Sequenz keine Positionen berücksichtigt, sondern einen „Bag-of-Word“ für die Sequenz bzw. die Merkmale im Fokus bildet. Diese Repräsentation kann zwar nicht mehr zwischen „Peter Müller“ und „Müller Peter“ unterscheiden, erfasst jedoch viele Gemeinsamkeiten zwischen „*Johann Peter Müller*“ und „*Peter Müller*“.

Darüber hinaus wurden zusätzliche Merkmale zur Abbildung der Anzahl Wörter der zu klassifizierenden Sequenz eingeführt: Zum einen war dies für jede beobachtete Länge ein binäres Merkmal, welches die Länge der Sequenz exakt abbildet; zum anderen ein Merkmal für die Sequenzlänge, dessen Wert durch

1/(Anzahl Wörter der Sequenz)

bestimmt wird.

Die erweiterte N-Gram Modellierung ist ein ausgesprochen innovativer Ansatz, der jedoch zusätzliche Heuristiken benötigt, um die Anzahl Instanzen nicht explodieren zu lassen. Wird

eine maximale Länge von beispielsweise fünf Wörtern für eine NE-Sequenz festgelegt, führt dies bei konsequenter Expansion jeder Sequenz zu einer Verfünffachung der Instanzen. Da der Zeitaufwand des gewählten Lernalgorithmus, der SVM, abhängig von der Anzahl der Instanzen ist, ist dieser Effekt äußerst unerwünscht. Allerdings kann dieser Effekt mit einfachen Beobachtungen und daraus abgeleiteten Heuristiken stark eingeschränkt werden:

- Eine NE-Sequenz überschreitet niemals eine Satzgrenze.
- Eine NE-Sequenz endet und beginnt nicht mit kleingeschriebenen Einheiten oder Satzzeichen.
- Anhand der Trainingsdaten können Wörter und Zweiwortsequenzen abgeleitet werden, die praktisch niemals Teil einer NE sind. Beispiele für Zweiwortsequenzen sind etwa „in den“ oder „sich die“. Diese Einträge stellen genau wie das Satzende Grenzen von NE-Sequenzen dar und reduzieren damit die Anzahl der zu klassifizierenden Instanzen. Ein solches Vorgehen zielt nicht darauf, ein umfangreiches Lexikon von nicht-NEs zu erstellen, sondern erzeugt durch einen hohen Frequenzschwellwert eine verhältnismäßig kleine Liste von Stoppwörtern und „Stopp-Sequenzen“.

Das skizzierte Auslassen von Einheiten kann dazu führen, dass Wörter und Sequenzen von der Klassifikation ausgeschlossen werden, die auch als NE vorkommen können. Weil eine manuelle bzw. intellektuelle Überprüfung schwer vorstellbar ist, kann dies nur anhand des Trainingskorpus getestet werden.

Da ein Wort Bestandteil mehrerer Sequenzen und damit Instanzen sein kann, kann es zu widersprüchlichen Klassifikationen kommen. Die Auflösung sich widersprechender Klassifikationsergebnisse wird in Abschnitt 5.4. beschrieben.

Beim automatischen Erwerb von Kontexten (5.3.2) wird bei der Modellierung des Kontexts die wortbasierte Repräsentation durch eine sequenzorientierte ersetzt. Die zu lernenden Kontexte können Wortsequenzen zwischen eins und drei Wörtern vor und nach den zu klassifizierenden Einheiten sein. Damit wird das oben festgelegte Fenster mit drei Wörtern davor und zwei dahinter zumindest für den Erwerb von Kontexten vergrößert, da nun auch hinter dem Fokus bis zu drei Wörter betrachtet werden.

5.2.2. Wortoberflächenmerkmale

Die wortweise Klassifizierung und die gewählte N-Gram Modellierung legt nur die zu klassifizierenden Instanzen und den betrachteten lokalen Kontext fest. Um Evidenz zur Identifizierung von NEs zu gewinnen, werden über alle Wörter im gewählten Ausschnitt

Informationen in Form von Oberflächenmerkmalen abgeleitet. Dazu werden die Wörter mit einer Reihe von regulären Ausdrücken getestet und einer der in Abbildung 5.3 aufgeführten Klasse zugewiesen. Diese deterministische Zuordnung zeigte in Vorstudien bessere Ergebnisse als Varianten, die eine Zugehörigkeit zu mehreren Klassen erlauben. Die Einteilung ist einigermaßen umfangreich und enthält auch Kategorien, über deren Qualität bzw. Berechtigung diskutiert werden kann. Allerdings sollte der Einfluss der spezielleren Klassen nicht überschätzt werden, schließlich setzt sich der Großteil der Tokens nur aus Buchstaben zusammen bzw. gehört zu den Satzzeichen. Im Trainingskorpus der deutschen CoNLL-Daten ([TJONG KIM SANG & DE MEULDER 2003]) sind dies 85%: 10% der Tokens bestehen aus Satzzeichen, 75% der Tokens nur aus Buchstaben; 48% sind kleingeschrieben, 26% fangen mit einem Großbuchstaben an und weniger als 1% bestehen nur aus Großbuchstaben. In anderen Korpora liegt eine ganz andere Verteilung vor.

Klassen von Wortoberflächenmerkmalen;		**Beispiel und Kommentare**
Das Kreuz in der mittleren Spalte zeigt an, dass eine eigene Klasse für diejenigen Wörter existiert, die mit einem Punkt enden.		
Satzendezeichen		*.?!;:*
Komma		*,*
Freistehender Bindestrich		*-*
Öffnende Klammer		*[({*
Schließende Klammer		*])}*
Anführungszeichen		*„ “*
Nur Großbuchstaben ohne Vokale	**x**	*FDP, WWW; BZW.*
Nur Großbuchstaben mit Vokalen	**x**	*CDU, FRANKFURT; ABZW.*
Einzelner Großbuchstabe	**x**	*F; P.*
Nur Großbuchstaben mit Bindestrich im Wortinnern		*MAIN-KINZIG-KREIS, S-VHS-C*
Mehrere Großbuchstaben mit mehreren Punkten		*E.T.A., F.D.P.*
Mehrere Buchstaben aber kein Vokal		*sd, pd* (Journalistenkürzel, Datenfehler)
Wort mit Groß- und Kleinbuchstaben im Wortinnern		*GmbH, StVO*
Großgeschrieben mit Bindestrich	**x**	*Versehrten-Sportgruppe, Leutheuser-Schnarrenberger; Josef-Str.*
Am Wortanfang großgeschriebenes Wort	**x**	*Maus, Frankfurt; Tel., Mio.*
Buchstaben mit Slash		*CDU/FDP*

Buchstaben mit Ziffern und optional Bindestrichen, Slash		*C3po, 90/Die* (Bestandteil von Bündnis 90/Die Grünen)
Folge von Kleinbuchstaben mit Vokalen und möglicherweise Bindestrich	**x**	*aus, der, verstimmt, nicht-öffentlich; bzw.*
Genau vier Ziffern		*1972, 1191* (möglicherweise Jahreszahlen)
Mehrere Ziffern, aber nicht genau vier	**x**	*3, 42, 10254; 121.*
Mehrere Ziffern mit Doppelpunkt dazwischen		*4:2, 72:76, 15:30* (Sportresultate, Uhrzeit)
Kombinationen von Ziffern mit weiteren Zeichen		*4,7, 12-18, 4/87*
Restgruppe, keines der erwähnten Muster		*B'90/Grüne, 5:1-Deckungssysteme, 1,9-Liter-Turbodiesel, 4#, a;ber,*

Abbildung 5.3: Übersicht über die verwendeten Wortoberflächenmerkmale

Auch wenn der Anteil spezieller Klassen nicht allzu groß ist, kann das Verfahren dennoch zur Ableitung Korpus-spezifischer Evidenz benutzt werden. Für biomedizinische NER etwa, kann das Muster „ATCG-Sequenz", welches eine DNA bezeichnet und dazu ihre Zusammensetzung aus den vier DNA-Bausteinen benutzt, effizient als weitere Klasse eingeführt werden.

Als weiteres Wortoberflächenmerkmal wurde auch die Wortlänge berücksichtigt. Dies ist ein äußerst triviales Merkmal, zeigte jedoch einen positiven Einfluss in den vorbereitenden Experimenten. Für die Merkmalskodierung der Wortlänge stehen grundsätzlich mehrere Möglichkeiten zur Verfügung. Beispielsweise könnte jeweils ein Merkmal für eine bestimmte Anzahl von Buchstaben verwendet werden. Allerdings sollte das Merkmal die Vermutung repräsentieren, dass sowohl ganz lange, als auch ganz kurze Wörter eher keine NEs sind und die erwähnte Kodierung ließe keine Bereiche der Wortlänge, sondern nur isolierte Kategorien von Wortlängen erkennen. Um sowohl für lange Wörter, wie auch für kurze Wörter ein Maß zu finden, wurde deshalb folgendes gewählt:

1/(Anzahl Buchstaben), welches bei kurzen Wörtern einen hohen Wert hat, und

1- 1/(Anzahl Buchstaben), welches bei langen Wörtern, einen hohen Wert annimmt.

5.2.3. Substring-Repräsentation von Wortformen

Die Merkmalsabbildung von Wortformen stellt eine interessante Herausforderung für lernbasierte Verfahren dar, schließlich ist die Anzahl der Wörter schier unendlich. Um spezifisches Wissen über Wörter zu repräsentieren, müsste also eine nahezu unendliche Menge spezifischer Merkmale gebildet werden. Nicht nur, dass das Lernen auf solchen

Mengen äußerst schwierig ist; es scheint auch nicht sinnvoll, für jedes Wort bzw. jede Wortform ein spezifisches Merkmal zu bilden. Gerade für die NER sind nur einige Wörter hilfreich, während andere uninteressant sind. Darüber hinaus enthält jedes Trainingskorpus, und sei es noch so gigantisch, immer nur eine begrenzte Anzahl Wörter, so dass das System auch in der Lage sein muss, mit unbekannten Wörtern umzugehen. Das Grammatikschreiben kann als manuelle Auswahl derjenigen Wörter gesehen werden, welche für die NER-Verfahren bedeutsam bzw. hilfreich sind. Wird auf eine solche manuelle Auswahl verzichtet, kann diese nur auf dem Trainingskorpus beruhen. [BORTHWICK 1999] etwa erzeugt für alle Wörter bzw. Wortformen, die öfter als dreimal im Korpus vorkommen, ein Merkmal. Aufgrund der reichen Morphologie des Deutschen, ist es fraglich, ob dies ein gangbarer Weg für deutsche Texte ist. Das CoNLL-03 Trainingskorpus für Englisch enthält knapp 24'000 Wortformen, während das deutsche Korpus bei identischem Umfang 33'000, also knapp 50% mehr Wortformen enthält. Darüber hinaus ist es fragwürdig, ob aufgrund der Zipf-Verteilung der Wortfrequenzen eine beliebig gesetzte Frequenz-Schwelle relevante von nicht hilfreichen Wörtern unterscheiden kann. Eine Lemmatisierung zur Reduktion der Wortformenmenge kommt aufgrund des strikt datenorientierten Ansatzes nicht in Betracht, obwohl eine Zusammenfassung verschiedener Flektionsformen eines Lexems möglicherweise wünschenswert ist.

Als Alternative wird deshalb eine Repräsentation mittels positionaler Substrings vorgeschlagen, die erstmalig in [RÖSSLER 2004b] beschrieben wurde. Die Idee dahinter ist, dass eine Wortform durch eine Menge ihrer Teilstrings repräsentiert wird. Dies führt zwar einerseits dazu, dass ein Wort nicht mit einem, sondern mit mehreren Merkmalen repräsentiert wird, andrerseits geschieht dies mit der Absicht, Ähnlichkeiten zwischen Wörtern zu erfassen. Dem Vorschlag liegt die Annahme zugrunde, dass für die NER in ähnlicher Weise Evidenz-beisteuernde Wörter einen ähnlichen Wortanfang oder ein ähnliches Wortende aufweisen. Dies gilt gleichermaßen für Komposita-, Flexions- und allgemeine morphologische Phänomene, wie beispielsweise:

- Komposita, die zur Erkennung von Personennamen hilfsreiche Funktionsbezeichnungen sind und alle denselben Kopf „*-sprecher*" enthalten: *Stadion-, Vorstands-, Firmen-, Schul-*
- Nachnamen, die alle denselben Beginn "*Schmi-*" aufweisen: *Schmidt, Schmid, Schmidheiny, Schmidthausen, Schmidtke.*
- Endungen, die eher auf Namen hinweisen oder eher dagegen sprechen. Die Suffix-Analyse hat sich bereits beim POS-Tagging bewährt [BRANTS 2000]. Beispielsweise

enden viele Nachnamen auf „*-er*“ (*Rössler, Hoeppner, Ziegler, Grünenfelder*) oder auf „*-ing*“ (*Schmelling, Gehring, Böhling*), manche Ortsnamen enthalten sehr ungewöhnliche Substrings wie „*-ydt*“ bei *Rheydt*, während andere Endungen typisch für gewöhliche Nomen sind „*-tion*“ in *Aktion, Demonstration.*

Die Repräsentation mit Substrings soll also in der Lage sein, Phänomene dieser Art abzudecken, aber gleichzeitig spezifische Informationen über das Wort bereitstellen. Die positionalen Substrings leisten dies in folgender Weise: Jede Wortform ergibt ein eindeutiges Muster aus den enthaltenen Substrings, während Ähnlichkeiten zwischen Wörtern über identische Substrings identifiziert werden. Die Integration der Position in das Substring-Merkmal führt dazu, dass zwei Wörter nur dann als ähnlich repräsentiert werden, wenn identische Substrings an derselben Position am Anfang oder am Ende des Wortes vorkommen.

Abbildung 5.4 zeigt die gewählte Repräsentation mit positionalen Substrings. In Vorexperimenten bewährte sich ein Maximum von acht positionalen Substrings, wobei sowohl Uni-, Bi-, vor allem aber Trigramme zur Repräsentation eines Wortes eingesetzt werden:

- Der letzte Buchstabe
- Die zwei letzten Buchstaben
- Ein Fenster von drei Buchstaben, das zwischen Wortende und Wortanfang hin und herwechselt, bis entweder sechs Buchstabentrigamme (drei von hinten, drei von vorne) extrahiert oder bis das Wort vollständig durch Trigramme abgebildet ist.

Position	Länge	Substrings
Wort: *Morphologie*		
Wortende	1	*e*
Wortende	2	*ie*
Wortende	3	*gie*
Wortanfang	3	*Mor*
Wortende-1	3	*ogi*
Wortanfang+1	3	*orp*
Wortende-2	3	*log*
Wortanfang+2	3	*rph*
Wort: *Hammer*		
Wortende	1	*r*
Wortende	2	*er*
Wortende	3	*mer*
Wortanfang+1	3	*Ham*
Wortende-1	3	*mme*
Wortanfang+2	3	*amm*

Abbildung 5.4: Exemplarische Repräsentation zweier Wörter mit positionalen Substring

Auf die Trainingsdaten des deutschen CoNLL-Korpus angewandt, führt dieses Verfahren zu etwa 15’000 Merkmalen vom Typ positionaler Substring. Würde hingegen jede Wortform des CoNLL-Korpus mit einem eigenen Merkmal repräsentiert, so erforderte dies 33’000 Merkmale.

Die Repräsentation mit Substrings kann auch in anderen Konfigurationen als den hier vorgeschlagenen durchgeführt werden: Mit und ohne positionale Angaben, mit Zwei-, Drei- oder Vier-Zeichenfenstern, mit einer vollständigen Reduktion auf Kleinbuchstaben und einer unterschiedlichen Maximalanzahl von Substrings. In Kapitel 6.1.1 sind Experimente mit unterschiedlichen Konfigurationen beschrieben, die das hier gewählte Verfahren mit anderen vergleichen.

Die vorgeschlagene Substring-Repräsentation wird sowohl auf das klassifizierende Wort als auch auf die vom N-Gram-Fenster abgedeckten Wörter des Kontexts angewandt. Bei der erweiterten N-Gram Modellierung ist die zu klassifizierende Sequenz nicht immer ein Wort, sondern kann auch aus mehreren Wörtern bestehen. Wie in 5.2.1 beschrieben wird bei der Erzeugung der Merkmale der zu klassifizierenden Sequenz die Position einzelner Wörter innerhalb der Sequenz nicht betrachtet, sondern es wird ein „Bag-of-Word“ Modell benutzt. Für die Substring-Repräsentation einer Sequenz wie beispielsweise „*Peter Müller*“ bedeutet dies: *„endet zweimal auf –er, startet einmal mit Pet-, startet einmal mit Mül-, etc.“* Diese

Repräsentation kann zwar nicht zwischen „Peter Müller“ und „Müller Peter“ unterscheiden, erfasst jedoch viele Gemeinsamkeiten zwischen „*Johann Peter Müller*“ und „*Peter Müller*“.

5.2.4. Überblick über die Basismerkmale und ihre Verwendung

Der Korpus-adaptive Ansatz verwendet als Basismerkmale zwei Gruppen von Merkmalen:

- Die Merkmale für die Wortoberfläche: Diese werden im Vorfeld festgelegt und können bei Bedarf erweitert werden. Zur NER in deutschen Texten wurden 31 Merkmale entwickelt. Unabhängig von der Anzahl der Merkmale erfüllt jedes Wort der Instanz immer nur eines der Merkmale. Darüber hinaus wird pro Wort der Instanz auch die Wortlänge mit zwei weiteren Merkmalen abgebildet.
- Merkmale über die in einem Wort enthaltenen positionalen Substrings: Nur die in den Trainingsinstanzen vorkommenden positionalen Substrings können vom Lernalgorithmus bewertet werden. Deshalb werden basierend auf der in 5.2.3 beschriebenen Konfiguration alle möglichen positionalen Substrings aus den Trainingsdaten extrahiert und als mögliche Merkmale bereitgestellt. Für die NER in deutschen Texten ergaben sich aus den Trainingsdaten des CoNLL-Korpus etwa 15’000 positionale Substrings. Von diesen 15’000 potentiellen Merkmalen erfüllt jedes Wort der Instanz aufgrund der in 5.2.3 beschriebenen Konfiguration jedoch maximal acht Merkmale.

Die Erzeugung der Basismerkmale kann formal durch die Merkmalsfunktion f notiert werden, die für die Erzeugung aller Merkmale zuständig ist. Die Merkmalsfunktion f überführt eine Instanz x, also ein zu klassifizierendes Wort und den umgebenden Kontext, in einen Merkmalsvektor $\vec{x}$:

$f : x \mapsto \vec{x}$, wobei $\vec{x}_i = f_i(x)$ und x_i die i-te Komponente von $\vec{x}$ ist und das Ergebnis der Merkmalsfunktion f_i ist.

Die meisten Basismerkmale sind binär, einzig die Merkmale zur Abbildung der Wortlänge sind numerische Werte. So überprüft beispielsweise die binäre Merkmalsfunktion f_k das Vorkommen eines bestimmten positionalen Substrings in einem Wort an einer bestimmten Position des N-Gram Fensters von x und ergibt in den meisten Fällen 0. Die ebenfalls binäre Merkmalsfunktion f_j testet, ob ein Wort an einer bestimmten Position von x nur aus

Großbuchstaben besteht. Einen numerischen Rückgabewert hat hingegen f_l, welche die Wortlänge kodiert.

$$f_j(x) = \begin{cases} 1 \text{ wenn (Wort\_im\_Fokus\_besteht\_nur\_aus\_Großbuchstaben } (x))=\text{true} \\ 0 \text{ sonst} \end{cases}$$

$$f_k(x) = \begin{cases} 1 \text{ wenn (Fokus+1\_endet\_auf\_“mer“}(x)) =\text{true} \\ 0 \text{ sonst} \end{cases}$$

$$f_l(x) = \begin{cases} 1/\text{Anzahl\_Buchstaben\_an\_Position\_Fokus-2 } (x) \end{cases}$$

Pro Wort an einer bestimmten Fensterposition ergeben sich maximal elf Basismerkmale mit einem Wert größer 0, nämlich ein Wortoberflächenmerkmal, zwei Wortlängenmerkmale und maximal acht Merkmale für positionale Substrings. Da diese Merkmale für alle sechs Wörter des gewählten Fensters erzeugt werden und die Angabe über die Position im Wortfenster enthalten, ergibt sich für die klassische N-Gram Modellierung ein Maximum von 66 Basismerkmalen pro Instanz. Diese stammen jedoch aus einem Merkmalsraum, der aufgrund der großen Anzahl positionaler Substrings sehr viel größer ist: Für das deutsche CoNLL-Korpus ergibt sich durch die etwa 15’000 möglichen Substring-Merkmale pro Wort des 6-Wort-Fensters ein Merkmalsraum von etwa 90’000 Merkmalen.

Die Abbildung 5.5 exemplifiziert die Repräsentation einer Instanz durch die Basismerkmale beim Einsatz der klassischen N-Gram Modellierung des Kontexts.

Satz: *Wie der Sanitäter Peter Hammer mitteilte, wird das Kind noch heute zurückkehren.*						
Wörter	der	Sanitäter	Peter	Hammer	mitteilte	,
	Position					
Merkmale	**-3**	**-2**	**-1**	**Fokus**	**+1**	**+2**
Deterministische Wortoberfläche	lowCase	upCase	upCase	upCase	lowCase	Komma
1/Wortlänge **1-(1/Wortlänge)**	0.33 0.67	0.11 0.89	0.2 0.8	0.16 0.84	0.11 0.89	1 0
Positionale Substrings	“r”_End ”er”_End ”der“_End	“r”_End ”er”_End ”S”_Start ”ter“_End “San“_Start “äte“_End-1 “ani“_Start+1 “tät“_End-2 “nit“_Start+2 “itä“_End-3	“r”_End ”er”_End ”P”_Start ”ter“_End “Pet“_Start “ete“_End-1	“r”_End ”er”_End ”H”_Start ”mer“_End “Ham“_Start ”mme“_End-1 ”amm”_Start+1	“e”_End ”te”_End ”m”_Start ”lte“_End “mit“_Start “ilt“_End-1 “itt“_Start+1 “eil”_End-2 “tte“_Start+2 “tei“_End-3	“,”_End

Abbildung 5.5: Exemplarische Merkmalsabbildung einer Instanz mit traditioneller N-Gram Modellierung

Bei der erweiterten N-Gram Modellierung, bei der nicht nur ein einzelnes Wort, sondern eine ganze Sequenz im Klassifikationsfokus steht, muss eine leicht modifizierte Abbildungsmethode verwendet werden. Da nun mehrere Wörter an der Position Fokus stehen, kann es beispielsweise vorkommen, dass ein positionaler Substring mehrmals an der Position Fokus vorkommt. Um dies abzubilden, sind die Merkmale der Position Fokus nicht mehr binär, sondern zählen, wie oft ein Merkmal erfüllt ist. Das Merkmal Wortlänge hingegen wird auf die durchschnittliche Länge der Wörter an der Position Fokus angewandt. Darüber hinaus wird ein weiteres Merkmal zur Abbildung der Anzahl der Wörter an der Position Fokus eingesetzt.

Satz: *Wie der Sanitäter Peter Hammer mitteilte, wird das Kind noch heute zurückkehren.*						
Wörter	wie	der	Sanitäter	Peter Hammer	mitteilte	,
	Position					
Merkmale	**-3**	**-2**	**-1**	**Fokus**	**+1**	**+2**
Deterministische Wortoberfläche	lowCase	upCase	upCase	upCase=2	lowCase	Komma
1/ ∅ Wortlänge **1-(1/ ∅ Wortlänge)**	0.33 0.67	0.33 0.67	0.11 0.89	0.18 0.82	0.11 0.89	1 0
Positionale Substrings	“e”_End ”ie”_End ”wie“_End	“r”_End ”er”_End ”der“_End	“r”_End ”er”_End ”S”_Start ”ter“_End “San“_Start “äte“_End-1 “ani“_Start+1 “tät“_End-2 “nit“_Start+2 “itä“_End-3	“r”_End=2 ”er”_End=2 ”H”_Start=1 ”mer“_End=1 “Ham“_Start=1 ”mme“_End-1=1 ”amm”_Start+1=1 ”P”_Start=1 ”ter“_End=1 “Pet“_Start=1 “ete“_End-1=1	“e”_End ”te”_End ”m”_Start ”lte“_End “mit“_Start “ilt“_End-1 “itt“_Start+1 “eil”_End-2 “tte“_Start+2 “tei“_End-3	“,”_End
1/ Länge der Sequenz im Klassifikationsfokus				0.5		

Abbildung 5.6: Exemplarische Merkmalsabbildung einer Instanz mit der erweiterten N-Gram Modellierung

5.3. Die Auswertung automatisch annotierter Daten

Alleine mit den in 5.2 beschriebenen Basismerkmalen ist kein leistungsfähiges NER-System vorstellbar. Was gänzlich fehlt, sind die in anderen Systemen vorhandenen bereits kategorisierten Listen von NEs oder von Triggern. In 4.4.3 ist die fehlende Korpus-Adaptivität solcher Listen ausführlich diskutiert worden. Um diese Schwierigkeit zu umgehen, verzichtet der hier entwickelte Ansatz gänzlich auf derartige Listen. Dennoch ist es notwendig, die Informationen, welche kategorisierte Listen bieten, zur Verfügung zu haben. Dazu werden Verfahren vorgeschlagen, welche diese Informationen durch eine Kombination aus nicht bzw. automatisch annotierten und annotierten Daten ableiten.

Zwei voneinander unabhängige Strategien werden zur Auswertung bzw. Nutzbarmachung von automatisch annotierten Daten eingesetzt. Die eine Strategie beruht auf der Ausnutzung der Struktur von Dokument- bzw. Texteinheiten (vgl. Abschnitt 3.2.6) in Verbindung mit bereits annotierten Einheiten. Diese sog. *Dokumenten-orientierte Klassifikation* kann als Standardverfahren der NER bezeichnet werden und wird in vielen Systemen, sowohl in regel- als auch lernbasierten Ansätzen eingesetzt. Das Verfahren wird als Bestandteil der Postprocessing-Komponente umgesetzt und wird im Rahmen des Systemüberblicks in 5.4.1 als Verarbeitungsschritt beschrieben. Die zweite Strategie hingegen ist ein neuer Ansatz, der

erstmalig in [RÖSSLER 2004b] beschrieben wurde und im Rahmen dieses Dissertationsprojekts entwickelt wurde. Der Kern des Ansatzes zielt auf die Ableitung von Evidenz für ein NER-Modell aus nicht-annotierten Daten und ist somit den in 4.3 vorgestellten Ansätzen zuzuordnen. Allerdings wird ein gänzlich neuer Weg vorgeschlagen, da anstelle der Extraktion zusätzlicher Instanzen für das Lernverfahren oder von Namens- oder Triggerlisten aus automatisch annotierten Daten neue Merkmale abgeleitet werden. Die automatischen Annotationen entstammen der Anwendung eines einfachen NER-Modells auf ein umfangreiches, nicht-annotiertes Korpus. Mit den neu gewonnenen Merkmalen werden die manuell annotierten Instanzen des Trainingskorpus angereichert. Auf den angereicherten Instanzen wird ein weiteres NER-Modell gelernt. Iterativ wird versucht, die gewonnenen Merkmale zu verbessern, indem das resultierende NER-Modell erneut auf das umfangreiche, nicht-annotierte Korpus angewendet wird, um anhand der besseren automatischen Annotationen genauere Merkmale abzuleiten. Das Verfahren ist in zwei unterschiedlichen Systemen umgesetzt und evaluiert worden: System I (5.3.1) beschränkt sich auf die Ableitung interner Evidenz. Die Weiterentwicklung des Ansatzes in System II (5.3.2) erwirbt simultan interne und externe Evidenz.

Sowohl System I und II als auch die Dokumenten-orientierte Klassifikation (5.4) werden durch die Idee einer *idealen Liste* mit kategorisierten NEs, NE-Bestandteilen oder Triggern verständlicher und nachvollziehbarer motiviert.

Ideale NE-Listen

Die Spezifikation der idealen NE-Listen ist erstmalig in [RÖSSLER 2004b] beschrieben worden. Die Idee ist ursprünglich auf Listen mit einzelnen Wörtern beschränkt, die möglicherweise Bestandteil einer NE sind oder auch isoliert als NE vorkommen können. Allerdings lassen sich wichtige, daraus gezogene Schlüsse auch auf Mehrworteinträge und auf externe Evidenz, also Trigger bzw. Wörter des NE-Kontextes übertragen.

Die idealen NE-Listen für eine NER-Anwendung sind abdeckend und verlässlich (vgl. dazu 5.1.2). Gäbe es keine mehrdeutigen Einträge, so genügte eine simple Aufzählung aller Wortformen, die zu einer der zu erkennenden NE-Kategorien gehören. Mehrdeutigkeit in Bezug auf die NE-Kategorie ist jedoch kein marginales Phänomen: Gemessen am deutschen Teil des CoNLL-Korpus sind 13% der Wortformen in 50’000 bzw. 24% in 200’000 Tokens mehrdeutig. Je größer ein Korpus ist, desto größer ist der Anteil der Wortformen, die mit mehr als einer NE-Kategorie vorkommen. Der reine Hinweis, dass

eine Wortform mehrdeutig ist, ist jedoch zu wenig spezifisch. Viel eher wird eine genaue Angabe zu dieser Mehrdeutigkeit benötigt.

	p(PER)	p(ORG)	p(LOC)	p(NIL)
Mark	0.15	0.05	0.1	0.7
Paris	0.04	0.1	0.85	0.01
Gulbuddin	1.0	0	0	0
Deshalb	0	0	0	1

Abbildung 5.7: Ausschnitt aus einer *idealen* NE-Liste

Eine Möglichkeit zur Repräsentation dieser Angabe ist, wie in Abbildung 5.7 gezeigt, die Wahrscheinlichkeit, dass eine Wortform w mit einer NE-Klasse c vorkommt, also

$p(c \mid w)$, wobei für die Beispiele in Abb. 5.7 $c_i \in$ {PER, ORG, LOC, NIL}, also auch immer die Klasse „keine-NE“ umfasst und für w gilt, dass

$$\sum_{i=1}^{n} p(c_i \mid w) = 1$$

Die ideale Liste ist wie erwähnt nur ein Konstrukt, dient aber der Reflexion über das Verhältnis von Wortformen zu NE-Klassen. Die gewählten Werte sind willkürlich und folgendermaßen zu interpretieren. „*Paris*“ etwa ist meist eine Ortsangabe, in manchen Fällen ein Personenname („*Paris Hilton*“), manchmal auch ein Organisationsname („*New Paris Theatre*“) oder keine NE der genannten Kategorien („*Festival du Film de Paris*“). Der Eintrag „*Mark*“ dagegen kommt meist als „Nicht-NE“ vor, was bei Texten möglich ist, in denen viele Geldbeträge in der älteren deutschen Währung „*Deutsche Mark*“ vorkommen. Gerade dieses Beispiel macht deutlich, was diese Werte eigentlich beschreiben und was ihr Gültigkeitsbereich ist. Diese Werte können als Abschätzungen verstanden werden, die anhand der NE-Frequenzen aus einem umfangreichen annotierten Korpus abgeleitet sind. Deshalb sind diese Frequenzen korpusspezifisch. Ein Zeitungskorpus aus dem letzten Jahr wird ein ganz anderes Verhältnis der NE-Vorkommen von „*Mark*“ aufweisen, da dies seit der Einführung des Euro keine aktuelle Währung bezeichnet. Die Korpusspezifizität dieser Werte macht das Korpus als eine wichtige Ebene zur Beschreibung und Beobachtung von

Wortformen in Bezug auf die NER-Aufgabe deutlich und führt zu klaren Kriterien für Listen von kategorisierten NE-Bestandteilen und Triggern: Derartige Listen müssen korpusspezifisch sein, wobei unter Korpus die zu verarbeitenden Texte zu verstehen sind und nicht etwa nur ein Trainings- oder Testkorpus. Darüber hinaus ist es erforderlich, die Mehrdeutigkeit der einzelnen Listeneinträge zu erfassen. Die Mehrdeutigkeit eines Eintrages ist spezifisch für ein zu verarbeitendes Korpus, d.h. diese ist von Korpus zu Korpus unterschiedlich. Es kann versucht werden, diese Mehrdeutigkeit mit einem numerischen Wert anzugeben, wobei ein hoher Wert für eine bestimmte Kategorie eine starke Tendenz bzw. Wahrscheinlichkeit anzeigt, dass der betreffende Eintrag mit der betreffenden Kategorie vorkommt.

Es sollte klar sein, dass zur Ableitung dieser numerischen Werte ein übliches Trainingskorpus absolut ungenügend ist. Um diese Werte abzuschätzen, sind von jedem Wort mehrere klassifizierte Vorkommnisse erforderlich. Darüber hinaus gewährleistet eine solche ideale Liste noch immer kein leistungsfähiges NER-System. Zwar könnten alle eindeutigen Wortformen problemlos klassifiziert werden, doch ist zur Auflösung der Mehrdeutigkeiten die Betrachtung des Wortes im Kontext erforderlich, also das konkrete Vorkommen. Ein konkretes Vorkommen einer Wortform ist zumindest für Menschen eindeutig, vorausgesetzt, dass definitorische Unschärfen und beabsichtigte Mehrdeutigkeiten, wie dies in satirischen Bemerkungen vorkommen kann, außer Acht gelassen werden. Das konkrete Vorkommnis zu klassifizieren, ist die eigentliche NER-Aufgabe. Die Klassifikation ist nicht in allen Fällen gleich schwierig. Einige Kontexte machen eine sehr klare Voraussage für eine bestimmte NE-Klasse (beispielsweise „*in <ORTSANGABE> ansässige*“). Solche Kontexte werden im Folgenden *prädiktive Kontexte* genannt. In Verbindung mit einer Dokumenten- bzw. Texteinheiten-orientierten Klassifikation können die NEs in prädiktiven Kontexten dazu benutzt werden, um NEs in nicht-prädiktiven Kontexten zu erkennen und zu klassifizieren (vgl. dazu Abschnitt 3.2.6).

Die NER als Klassifikation aller Wörter eines Textes kann also auf Evidenz aus zwei unterschiedlichen Betrachtungsweisen zurückgreifen:

1. Listen von NEs, welche durch prädiktive Kontexte erkannt wurden und spezifisch für ein Dokument bzw. eine Texteinheit sind. Dies ist ein Standardverfahren der NER und wird im Rahmen dieses Systems als Postprocessing-Verarbeitungsschritt „Dokumenten-orientierte Klassifikation“ im Abschnitt 5.4 beschrieben.

2. Listen von NEs, NE-Bestandteilen und Triggern, welche für ein bestimmtes Korpus spezifisch sind. Abschnitte 5.3.1 und 5.3.2 stellen zwei Varianten eines selbst entwickelten Verfahrens vor, welches die Evidenz, die in solchen Listen enthalten ist, durch eine Kombination aus manuell annotierten Daten und großen Mengen des zu verarbeitenden Korpus automatisch ableitet.

Grundlegende Herangehensweise beim Einsatz nicht-annotierter Daten

Einzig auf die in 5.2 beschriebenen Basismerkmale beschränkt, fehlen dem NER-System die in anderen Systemen vorhandenen, bereits kategorisierten Listen von NEs oder von Triggern. In diesem Abschnitt wird die prinzipielle Herangehensweise beschrieben, mit der die durch solche Listen bereitgestellte Evidenz durch eine Kombination aus nicht bzw. automatisch annotierten und annotierten Daten abgeleitet wird. Im Gegensatz zu den in 4.3 vorgestellten Ansätzen werden keine neuen Trainingsinstanzen erzeugt, sondern es treten die erzeugten Ressourcen in der Form zusätzlicher, sog. *erweiterter Merkmale* in Erscheinung. Diese werden zur Anreicherung der Repräsentation der Instanzen eingesetzt. Die Merkmale werden aus automatisch annotierten Daten gewonnen und unterstützen den Lern- und Klassifikationsprozess.

Abbildung 5.8 zeigt einen Überblick über die erforderlichen Schritte zur Erzeugung der erweiterten Merkmale. Die Abläufe sind sowohl für das System I (5.3.1) zum Erwerb von interner Evidenz als auch für den gleichzeitigen Erwerb von interner und externer Evidenz in System II (5.3.2) identisch.

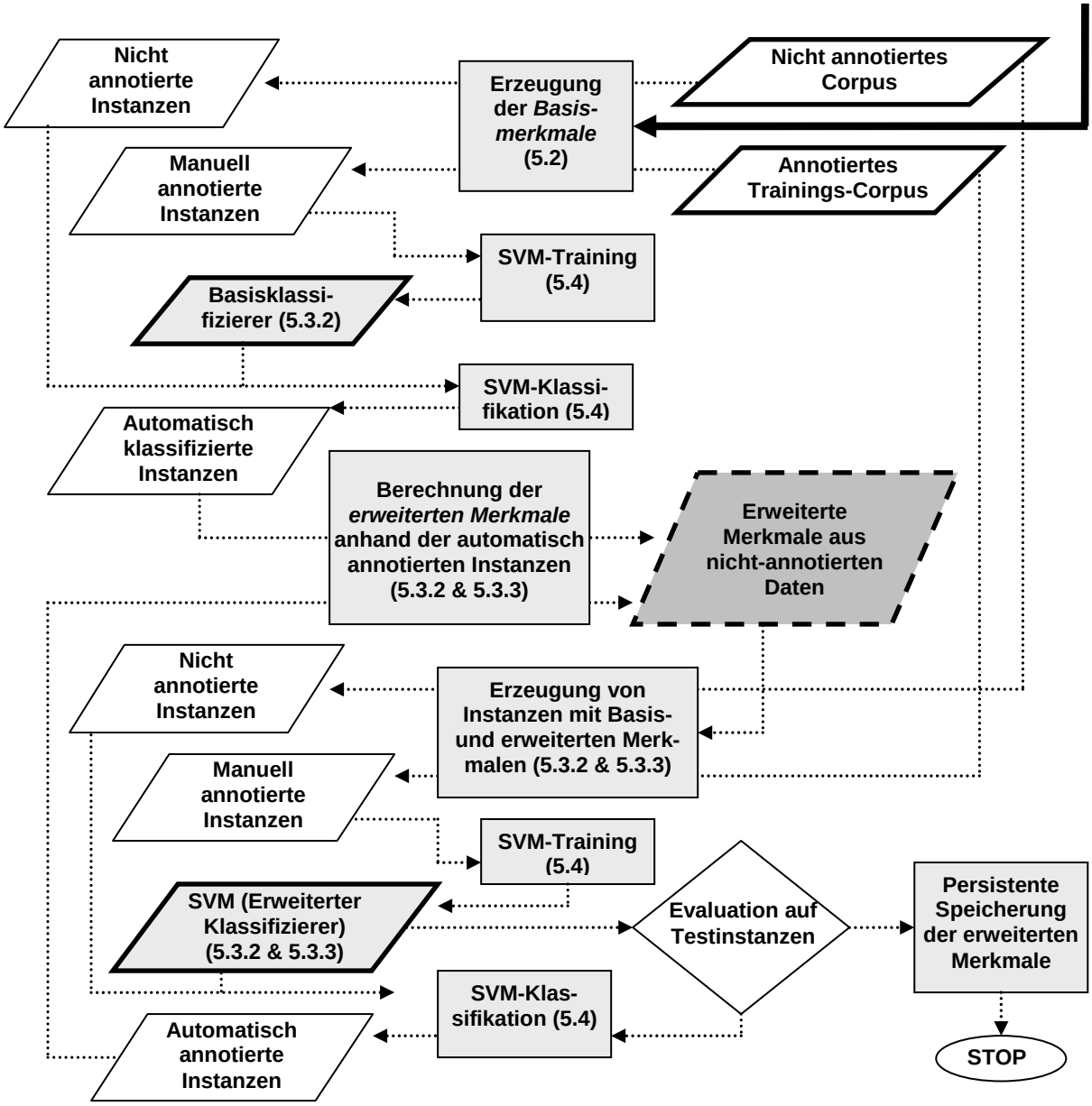


Abbildung 5.8: Architektur zur Erzeugung der erweiterten Merkmale aus automatisch annotierten Daten

Zentral für den Ansatz ist die Kombination aus annotierten und nicht-annotierten Daten. Die annotierten Daten in Form eines Trainingskorpus dienen dazu, ein initiales NER-Modell zu erwerben, den sog. *Basisklassifizierer*. Der Basisklassifizierer wird mit mehreren SVMs realisiert, die jeweils für eine NE-Klasse auf den in 5.2 beschriebenen Basismerkmalen der Trainingsinstanzen trainiert werden. Mit dem Basisklassifizierer, der für Precision und nicht für Recall optimiert sein sollte, werden große Textmengen annotiert. Diese nicht-annotierten Daten bestehen aus einem umfangreichen Korpus der zu verarbeitenden Texte. Die durch den

Basisklassifizierer erzeugten Annotationen stellen die Grundlage zur Erzeugung der erweiterten Merkmale dar. Dabei werden für alle Wortformen, die oft genug im Korpus vorkommen, auf den automatischen Annotationen basierend Merkmale berechnet, welche durch die Idee der idealen NE-Liste (vgl. oben) inspiriert sind. Um beispielsweise interne Evidenz bereitzustellen, enthalten diese zusätzlichen Merkmale Information darüber, wie die Vorkommen einer bestimmten Wortform in einem großen Korpus von dem einfachen NER-Modell klassifiziert wurden. Um externe Evidenz bereitzustellen, ist ein Maß gesucht, welches in den automatisch annotierten Texten diejenigen Wörter bzw. Wortsequenzen hoch gewichtet, welche oft vor bzw. hinter einer bestimmten NE-Klasse zu beobachten sind und deswegen als Trigger fungieren können. Die so berechneten, erweiterten Merkmale stehen anschließend für alle Wortformen zur Verfügung, die oft genug im Korpus vorkamen. Nach der Anreicherung der Trainingsinstanzen durch diese erweiterten Merkmale wird ein neuer Klassifizierer, der *erweiterte Klassifizierer*, gelernt. Durch den Lernvorgang werden die erweiterten Merkmale im Modell berücksichtigt. Damit wird es dem Lernalgorithmus überlassen, die neu gewonnenen Merkmale zu bewerten und mit Gewichten im Modell zu berücksichtigen. Dies federt den Einfluss von Fehlern ab, die aufgrund von falschen Annotationen in die erweiterten Merkmale gelangen. Der erweiterte Klassifizierer wird anhand von Testinstanzen evaluiert. Führen die neu erworbenen Merkmale zu einem besseren Modell, so kann der Vorgang wiederholt werden. Ist die Erkennungsrate besser als der vorhergehende Klassifizierer, was im ersten Durchgang der Basisklassifizierer ist, wird der erweiterte Klassifizierer erneut auf das umfangreiche Korpus angewandt und die Werte der erweiterten Merkmale werden erneut berechnet. Dieser Vorgang kann solange wiederholt werden, bis keine Verbesserung in der Erkennungsleistung mehr festzustellen ist. Ist dies der Fall, so werden die erweiterten Merkmale persistent gespeichert und der Prozess endet.

Der erweiterte Klassifizierer, basierend auf den Basismerkmalen und den erweiterten Merkmalen des letzten Durchgangs ist das Modell, welches beim eigentlichen Einsatz des NER-Systems benutzt wird. Die Ausgabe dieses Modells kann anschließend durch weitere Komponenten, wie beispielsweise die Dokumenten-orientierte Klassifikation (5.4) optimiert werden.

Wird ein solches Verfahren auf das gesamte oder einen Großteil des eigentlich zu verarbeitenden Korpus angewandt, so sind die erzeugten Ressourcen in der Form der erweiterten Merkmale zum einen weitestgehend abdeckend, zum anderen spezifisch. Die

Abdeckung erstreckt sich aber nicht auf Einheiten, die nur wenige Male im Korpus vorkommen.

5.3.1. Interne Evidenz aus automatisch annotierten Daten (System I)

Da zu Beginn keine erweiterten Merkmale zur Verfügung stehen, beruht der initiale Basisklassifizierer nur auf dem Modell, das anhand der Basismerkmale der annotierten Daten gelernt wurde. Da der Basisklassifizierer nicht auf externe Evidenz beschränkt ist, ist auch interne Evidenz im Modell enthalten, in der Form derjenigen NEs, die im Trainingkorpus vorkommen. Da der Basisklassifizierer in seiner Erkennungsleistung beschränkt ist und im besten Fall eine hohe Precision, aber einen schwächeren Recall aufweist, entsprechen die gewonnenen Merkmale aufgrund der schwachen Ausbeute ganz und gar nicht den gewünschten Werten der idealen NE-Liste. Das heißt, die Klasse „keine-NE“ ist deutlich überrepräsentiert. Allerdings besteht durchaus die Hoffnung, dass sich aus den erzeugten Werten Tendenzen über die NE-Zugehörigkeit der jeweiligen Wortformen ablesen lassen. Diese Hoffnung begründet sich auf der Annahme, dass der eingesetzte Klassifizierer vor allem NEs in prädiktiven Kontexten erkennt und alle NEs eine ähnliche Tendenz haben, in einem prädiktiven Kontext vorzukommen.

Um dies zu gewährleisten, könnte ein Klassifizierer eingesetzt werden, der nur auf die Merkmale der externen Evidenz Zugriff hat. Allerdings hat sich bereits in Vorexperimenten mit den deutschen CoNLL-Daten gezeigt, dass ein solcher Klassifizierer eine derartig schwache Erkennungsleistung aufweist, dass ein darauf aufbauender Erwerb von neuen Merkmalen zur internen Evidenz für deutsche Texte undenkbar ist.

Deshalb wurde auf eine Aufteilung der Evidenzen auf mehrere sich ergänzende Lernaufgaben verzichtet. Darüber hinaus führt die wortbasierte N-Gram Modellierung dazu, dass bei einer NE, die aus mehreren Wörtern besteht, nicht nur der eigentliche Kontext, sondern auch Bestandteile der NE als Kontext im N-Gram Fenster auftauchen. Die nicht-distinktive Großschreibung bzw. das daraus resultierende Phänomen der unklaren Grenzen einer NE-Sequenz (vgl. dazu Kap. 3.2.5) führt zu einem Verwischen zwischen interner und externer Evidenz: In vielen Fällen ist aufgrund der Merkmale nicht zu entscheiden, ob ein Wort direkt vor oder nach dem Fokuswort Teil einer möglichen NE ist oder aber zum Kontext gehört. Inwiefern sich das positiv oder negativ auf die Lernbarkeit auswirkt, muss in Experimenten untersucht werden.

Für die Auswertung automatisch annotierter Daten bzw. die daraus resultierenden Merkmale wird ein Verfahren benötigt, welches folgenden Anforderungen genügt:

- Das Verfahren muss in der Lage sein, die Mehrdeutigkeit von Wortformen („Paris“ als PERSON und LOCATION etc.) zu berücksichtigen.
- Es muss tolerant gegenüber möglichen Falschklassifikationen des Basisklassifizierers sein.
- Es muss damit umgehen können, dass die Grenzen zwischen interner und externer Evidenz nicht immer eindeutig sind.

Orientiert man sich an der idealen NE-Liste, so bietet sich die relative Häufigkeit r eines Wortes w in Verbindung mit einer NE-Klasse c an:

$$r(w,c) = \frac{freq(w,c)}{freq(w)}$$

Diese Abbildung ist in der Lage, Mehrdeutigkeiten zu repräsentieren. Käme also das Wort „Zürich“ viermal als Ortsangabe, zweimal als Organisation und vierzehnmal als keine-NE vor, so ergäbe sich

r(Zürich, LOC)=0,2; r(Zürich, ORG)=0,1; r(Zürich, NIL)=0,7; r(Zürich,PER)=0

Diese Werte können direkt als Merkmal für semantische Eigenschaften des Wortes „*Zürich*“ verwendet werden und zur Erweiterung des Basisklassifizierers benutzt werden. Dazu werden die Trainingsinstanzen durch die gewonnenen erweiterten Merkmale angereichert. Auf diesen angereicherten Instanzen wird erneut ein Modell trainiert, welches den Basisklassifizierer ersetzt. Wenn die erweiterten Merkmale Evidenz bereitstellen, wird der erweiterte Klassifizierer den Basisklassifizierer an Erkennungsleistung übertreffen. Die erweiterten Merkmale basieren also auf der Klassifikation von unterschiedlichen Vorkommen einer sprachlichen Einheit und versuchen daraus ein Merkmal abzuleiten, welches nicht nur für die betrachtete sprachliche Einheit Gültigkeit hat, sondern die „NE-Haftigkeit“ aller vorkommenden sprachlichen Einheiten kodiert. Der erweiterte Klassifizierer wird nun iterativ auf das nicht-annotierte Korpus angewandt, um anhand der Annotationen die semantischen

Merkmale neu zu berechnen. Das Verfahren endet, wenn keine Verbesserung mehr erreicht werden kann.

Da die berechneten Werte nicht Wahrscheinlichkeiten, sondern anhand der Trainingsdaten gewichtete Merkmale darstellen, spielt es auch keine Rolle, dass der Wert der „keine-NE“ Klasse deutlich überhöht ist. Um die manchmal unscharfen Grenzen zwischen interner und externer Evidenz zu berücksichtigen, kann das automatisch extrahierte Merkmal nicht nur für das Fokuswort benutzt werden, sondern auch für die Wörter im Kontext.

Die geschilderte Anreicherung von Instanzen mit *zusätzlichen Merkmalen* steht im Gegensatz zu den in 4.3 geschilderten Verfahren, in denen *zusätzliche Instanzen* oder NE-Listen erworben werden. Die hier verfolgte Strategie weist zwei äußerst interessante Eigenschaften auf. Die Anzahl der Instanzen bleibt trotz der Berücksichtigung nicht-annotierter Daten konstant. Das erst ermöglicht den Einsatz der SVM als Lernverfahren. Schließlich verhält sich der Trainingsaufwand der SVM superlinear zur Anzahl der Instanzen, so dass eine Erweiterung der Anzahl der Instanzen nur mit einem deutlich höheren Zeitaufwand für das Training möglich ist. Die zweite interessante Eigenschaft beruht darauf, dass die aus den nicht-annotierten Daten gewonnene Evidenz in Form von Merkmalen integriert und deshalb von der SVM gewichtet wird. Da der Einsatz nicht-annotierter Daten fast zwangsläufig Fehler erwirbt, kann dies durch die SVM gewissermaßen abgefedert werden.

Die Idee, interne Evidenz aus automatisch annotierten Daten zur Merkmalsanreicherung zu gewinnen, wurde an einer Vielzahl von Versuchen mit dem Basisklassifizierer und seiner Anwendung auf die 40 Millionen Wörter des FR-Korpus ([FR-CORPUS 1994]) geschärft und weiterentwickelt. Ausgehend von der Idee der idealen Liste (vgl. Abschnitt 5.2.6) wurden mit der Kategorie PERSON zahlreiche Varianten des Verfahrens entwickelt und überprüft. Um die gewählte Umsetzung, welche auch in [RÖSSLER 2004b] beschrieben ist, zu verstehen, wird auch eine Vorläuferversion beschrieben. Diese Vorläuferversion, die naive Umsetzung der Idee der idealen Liste, kann gewissermaßen als Ausgangspunkt betrachtet werden und wird zuerst vorgestellt. Die Evaluation der beiden Verfahren wird in 6.3.1 beschrieben. Die erweiterten N-Gramme (Abschnitt 5.2.1) sind erst nach der Umsetzung dieser Verfahren entwickelt worden, so dass die in diesem Abschnitt beschriebenen Systeme auf der einfachen N-Gram Modellierung des Kontextes beruhen.

Der naive Ansatz beruht auf der sehr direkten Umsetzung der Idee der idealen Liste, in dem für jede Wortform v, welche in den automatisch annotierten Daten vorkommt, aufgrund der

Zuweisung der NE-Klasse y des Basisklassifizierers ein Merkmal mit dem folgenden Wert erzeugt wird:

$$\frac{freq(v,y)}{freq(v)}$$

Kommt also eine Wortform wie „Richter“ insgesamt 100-mal vor und wird davon 20-mal als PERSON annotiert, so wird für diese Wortform ein Merkmal mit dem Wert 0.2 erzeugt. Das Merkmal bildet gewissermaßen die Klassifikationsgeschichte jeder Wortform bzw. all seiner Vorkommen in den automatisch annotierten Daten ab. Bei dem naiven Ansatz entsteht so pro Klassifizierer bzw. NE-Klasse ein Merkmal. Dieses Merkmal wird bei der Erzeugung neuer Trainingsdaten für den erweiterten Klassifizierer bei jedem Vorkommen der Wortform „Richter“, sei es im Fokus oder im Klassifikationskontext, mit dem berechneten Wert für die Instanz hinzugefügt. Diese Werte werden individuell für alle Wortformen berechnet und als Merkmal eingesetzt, wenn diese häufiger als n-mal durch den Basisklassifizierer annotiert wurden. Allerdings zeigte sich in Experimenten (Abschnitt 6.3.1), dass das naive Verfahren den erweiterten Klassifizier zwar optimierte, aber in viel zu geringem Maße.

Die anschließend gewählte Umsetzung basiert auf der Idee, den Klassifizierentscheid detaillierter abzubilden. Das naive Verfahren geht von einem rein binären Klassifikationsentscheid aus, obwohl die Klassifikation auf sehr unterschiedlichen Evidenzen beruhen kann. Handelt es sich um ein Wort, welches bereits im Trainingskorpus vorkam und in einem ebenfalls aus dem Trainingskorpus bekannten Kontext vorkommt, so ist die Klassifikation sehr viel verlässlicher, als wenn etwa ein unbekanntes Wort in einem unbekannten Kontext vorrangig aufgrund seiner Endung klassifiziert wird. Allerdings ist die Ausgabe der SVM (Abschnitt 4.2.5), welche als Lernalgorithmus eingesetzt wurde, nicht direkt als Wahrscheinlichkeit interpretierbar. Die Abschätzung von Wahrscheinlichkeiten mit SVMs (vgl. etwa [RÜPING 2004]) ist keine triviale Herausforderung.

Um dennoch eine feinere Abbildung des Klassifikationsentscheids zu haben, wird eine Diskretisierung des Wertes von $(\vec{w}\cdot\vec{x}+b)$ vorgeschlagen. Die klassische SVM-Entscheidungsfunktion klassifiziert Instanzen basierend auf $sign(\vec{w}\cdot\vec{x}+b)$, so dass positive Werte als zur Klasse gehörend und negative als nicht zur Klasse gehörend interpretiert werden. Die Diskretisierung erweitert diese Entscheidungsfunktion gewissermaßen, indem - zumindest für die Merkmalsberechnung - unterschieden wird, auf

welchem Wert eine Klassifikation beruht. Aufgrund der Diskretisierung wird für jede Wortform v, welche mindestens zweimal mit einem Wert in einem bestimmten Wertebereich für die NE-Klasse y vorkam, ein Merkmal mit dem folgenden Wert erzeugt:

$$\frac{freq(v, \theta_1 > (\vec{w} \cdot \vec{x} + b) > \theta_2)}{freq(v)}$$

wobei für die Schwellenwerte θ die Reihe {>1, >0.5, >0, >-0.5, >-1, <=-1} von links nach rechts durchlaufen wird. Diese Werte sind intuitiv festgelegt worden, haben sich jedoch experimentell bewährt.

Zusätzlich wird aus der absoluten Frequenz, mit der eine Wortform v mit der NE Klasse y vorkam, ein Merkmal mit dem folgenden Wert erzeugt:

$$\log(freq(v, y) + 1)$$

Die Berücksichtigung der absoluten Frequenz gibt einen Hinweis darauf, auf wie vielen Vorkommen die berechneten Werte beruhen und könnte damit als eine Art Verlässlichkeitsmaß für die berechneten Werte interpretiert werden.

Damit werden pro NE-Klasse maximal sieben Merkmale erzeugt. Da bei der Instanzenanreicherung jedoch die erweiterten Merkmale aller NE-Klassen hinzugefügt werden, ergibt dies bei PERSON, ORGANISATION und LOCATION 21 Merkmale bzw. angewandt auf alle sechs Wörter des Wortfensters 186 potentielle Merkmale. Grundsätzlich könnten für jede Klassifikationsaufgabe nur diejenigen erweiterten Merkmale benutzt werden, die der aktuellen Klassifikationsaufgabe entsprechen. Hier wird allerdings ein anderer Weg beschritten: Es werden für alle Klassifikationsaufgaben dieselben Merkmale benutzt. Damit hat der Klassifizierer für PERSON gewissermaßen auch Zugriff auf die erweiterten Merkmale für LOCATION, die beispielsweise aussagen, dass die zu klassifizierende Einheit oft als LOCATION klassifiziert wurde. Dieses Merkmal kann auch für die Klassifikationsaufgabe PERSON hilfreich sein. Darüber hinaus ist es mit erheblich weniger Aufwand verbunden, wenn die unterschiedlichen SVMs auf dieselben Daten angewandt werden können und nicht jeweils eine neue Repräsentation erzeugt werden muss. Zwar erfordert das Training weiterhin

unterschiedliche Daten, doch unterscheiden sich diese nur hinsichtlich der Label, nicht aber hinsichtlich der Repräsentation der Instanzen.

Abbildung 5.9 zeigt die erweiterten Merkmale an einem Beispiel. Die eingetragenen Werte sind nur zur Verdeutlichung gewählt worden. Das Wort „Peter“ etwa, hat eine Tendenz, als PERSON klassifiziert zu werden und kommt offensichtlich auch mehrmals als ORGANISATION bzw. als Bestandteil davon vor. Die kleingeschriebenen Einheiten haben keine erweiterten Merkmale, da sie nicht im Fokus einer Instanz auftraten. Die vielen Nullen bei den Wörtern mit erweiterten Merkmalen kommen daher, dass ein Merkmal nur mit einem Wert versehen wird, wenn der Wert von $(\vec{w}\cdot\vec{x}+b)$ zweimal oder häufiger in einem Wertebereich lag. Deshalb ergeben die summierten Werte auch nicht eins.

Satz: *Wie der Sanitäter Peter Hammer mitteilte, wird das Kind noch heute zurückkehren.*							
Klassifizierer		wie	der	Sanitäter	Peter	Hammer	mitteilte
	ERWEITERTE MERKMALE	**Position**					
	Frequenz von $\vec{w}\cdot\vec{x}+b$ im Wertebereich:	**-3**	**-2**	**-1**	**Fokus**	**+1**	**+2**
PER	**>1**	-	-	0	0.2	0	-
	<=1 && >0.5	-	-	0	0.1	0	-
	<=0.5 && >0	-	-	0.05	0.1	0.1	-
	<=0 && >-0.5	-	-	0	0.1	0.2	-
	<=-0.5 && >-1	-	-	0	0.1	0.1	-
	<=-1	-	-	0.9	0	0.5	-
ORG	**>1**	-	-	0	0	0	-
	<=1 && >0.5	-	-	0.1	0	0.05	-
	<=0.5 && >0	-	-	0.2	0.1	0.1	-
	<=0 && >-0.5	-	-	0.1	0.1	0.05	-
	<=-0.5 && >-1	-	-	0	0.2	0	-
	<=-1	-	-	0.5	0.5	0.7	-
LOC	**>1**	-	-	0	0	0	-
	<=1 && >0.5	-	-	0	0	0	-
	<=0.5 && >0	-	-	0	0	0	-
	<=0 && >-0.5	-	-	0	0	0	-
	<=-0.5 && >-1	-	-	0	0	0	-
	<=-1	-	-	0.9	0.9	0.9	-
log(*freq*(*v*,*y*)+1) **PER**		-	-	0.3	1.3	0.48	-
log(*freq*(*v*,*y*)+1) **ORG**		-	-	0.6	0.48	0.48	-
log(*freq*(*v*,*y*)+1) **LOC**		-	-	0	0	0	-

Abbildung 5.9: Die erweiterten Merkmalen der internen Evidenz mit exemplarischen Werten

5.3.2. Der simultane Erwerb von interner und externer Evidenz (System II)

Das in 5.3.1 beschriebe System I zeigt äußerst ermutigende Resultate, die in [RÖSSLER 2004b] veröffentlicht wurden. Allerdings hat das Verfahren zwei wichtige Beschränkungen. Zum einen extrahiert der Ansatz nur interne Evidenz, der Erwerb neuer Kontexte findet nicht statt. Zum anderen wird durchgehend die wortweise Betrachtung und Verarbeitung angewendet; es werden also keine NE-Sequenzen gelernt. Um diese Beschränkungen anzugehen, wird das Verfahren für das System II folgendermaßen modifiziert:

- Um die Beschränkung auf die interne Evidenz anzugehen, wird simultan interne und externe Evidenz erworben, d.h. es werden aus den automatisch annotierten Texten Kontexte von NEs gelernt.
- Die aus den automatischen Annotationen abgeleiteten Merkmale beschränkten sich bisher auf Sequenzen der Länge eins. Diese Begrenzung wird aufgehoben, d.h. es werden auch Merkmale über Mehrworteinheiten erzeugt.
- Der Einsatz der in 5.2.2 beschriebenen erweiterten N-Gram Modellierung ermöglicht eine stärkere Beachtung des Mehrwortcharakters von NEs und ist möglicherweise geeignet, die bisherige wortweise Betrachtung und Verarbeitung zu ersetzen oder zu ergänzen. Darüber hinaus zeigt die Evaluierung des Verfahrens als Basisklassifizierer in 6.1.5 eine höhere Precision als die traditionelle N-Gram Modellierung. Damit eignet sich der auf den erweiterten N-Grammen beruhende Basisklassifizierer prinzipiell besser für den Erwerb von Evidenz aus automatischen Annotationen.

Davon abgesehen entsprechen die erforderlichen Abläufe genau der Erzeugung der erweiterten Merkmale der in Abb. 5.8 skizzierten Architektur.

Externe Evidenz und Korpus-Adaptivität

Im Rahmen eines Korpus-adaptiven Ansatzes ist es vor den Ausführungen des eigentlichen Verfahrens erforderlich, einige Überlegungen zur Korpusspezifizität und zur Voraussagekraft externer Evidenz anzustellen. Für prädiktive Kontexte kann wie für die meisten Verteilungen sprachlicher Phänomene eine Zipf-Verteilung angenommen werden. Das bedeutet, einige wenige kommen ganz oft vor und sind deshalb auch auf unterschiedlichen Korpora hilfreich für die NER. Der Großteil der Kontexte hat jedoch eine geringe Frequenz und es ist zu vermuten, dass viele davon in bestimmten Korpora häufiger auftreten als in anderen. So kommt etwa im Sportteil von deutschen Tageszeitungen der hilfreiche Kontext *„Bundesligist“*

direkt vor dem Namen einer Fußballmannschaft vor. Außerhalb des Sportteils wird dieses Wort hingegen kaum verwendet. Die Zipf-Verteilung bedeutet, dass ein Trainingskorpus hilfreiche Kontexte nur zu einem Teil enthält und ein automatischer Erwerb hilfreicher Kontexte erstrebenswert ist. Werden Kontexte erworben und repräsentiert, müssen auch Möglichkeiten gefunden werden, die Voraussagekraft eines Kontextes zu repräsentieren. Diese Voraussagekraft beschreibt, dass ein Kontext wie „*die Firma XYZ produziert*“ eine sehr viel verlässlichere Voraussage über die NE-Zugehörigkeit von *XYZ* macht als etwa „ist in *XYZ*.“. Letzteres könnte eine Ortsangabe, aber auch ein gewöhnliches Nomen („*Sorge*“, „*Aufruhr*“) sein. Gerade beim automatischen Erwerb externer Evidenz wird deutlich, dass die Vorstellung vom verlässlichen Trigger für eine bestimmte NE-Kategorie nicht aufrecht erhalten werden kann. Bedingt durch die freie Wortstellung des Deutschen ist etwa der Kontext „*hat XYZ angekündigt*“ mehrdeutig, da *XYZ* sowohl eine Person, eine Organisation oder ein gewöhnliches Nomen (z.B. „Entlassungen“) sein kann.

Kontextmerkmale zur Repräsentation von externer Evidenz

Für die Kontextmerkmale zur Abbildung externer Evidenz wird eine Repräsentation gewählt, die sich an der idealen Liste in Abbildung 5.7 orientiert. Die Berechnung des Merkmals orientiert sich an dem oben beschriebenen Erwerb interner Evidenz aus automatisch annotierten Daten. Die Grundlage der Merkmalsberechnung ist die Beobachtung, wie oft ein bestimmter Kontext vor oder nach einer bestimmten NE-Klasse vorkommt.

Das Kontextmerkmal dient der Ableitung von Wissen über NE-Kontexte. Alle Sequenzen der Länge eins bis drei Wörter bzw. Einheiten, die unmittelbar vor oder nach einer NE vorkommen, sind potentielle Kontexte. Die Unterscheidung nach Länge ({1,2,3}) und Position zur NE {vor, nach} ergibt sechs unterschiedliche Kontexttypen pro NE-Klasse.

Abbildung 5.10 zeigt, wie diese Wahl der Kontextausschnitte bei drei NE-Klassen zu 18 Kontexttypen führt. Für jeden Kontexttyp werden zwei Merkmale erzeugt. Das eine Merkmal basiert auf der relativen, das andere auf der absoluten Frequenz.

„*a b c Fokus d e f*“ als abstrakte Form einer zu klassifizierenden Einheit im Fokus mit drei Wörtern davor und dahinter.								
Satz		**a**	**b**	**c**	**Fokus**	**d**	**e**	**f**
Position		**-3**	**-2**	**-1**		**+1**	**+2**	**+3**
Mögliche Kontexte aus einem („a“) bis drei Wörtern („a b c“) vor und hinter einem Fokus	**PER**			a				
						d		
			a	b				
						d	e	
		a	b	c				
						d	e	f
Mögliche Kontexte aus einem („a“) bis drei Wörtern („a b c“) vor und hinter einem Fokus	**ORG**			a				
						d		
			a	b				
						d	e	
		a	b	c				
						d	e	f
Mögliche Kontexte aus einem („a“) bis drei Wörtern („a b c“) vor und hinter einem Fokus	**LOC**			a				
						d		
			a	b				
						d	e	
		a	b	c				
						d	e	f

Abbildung 5.10: Als Grundlage der Kontextmerkmale gewählte Kontextausschnitte

Für jede Sequenz s, die einer der in 5.10 gezeigten Kontextausschnitten entspricht und die mindestens einmal als NE-Kontext einer Kategorie y vorkommt, wird ein Kontextmerkmal mit der relativen Frequenz berechnet:

$$\frac{freq(s,y)}{freq(s)}$$

Hierbei wird vorausgesetzt, dass der resultierende Wert größer als 0.01 ist. Dieser Schwellwert eliminiert Sequenzen, die mit hoher Frequenz im Korpus vorkommen und dennoch manchmal als NE-Kontext vorkommen. Neben der relativen Frequenz wird für jede Sequenz ein weiteres Merkmal zur Abbildung der absoluten Frequenz berechnet:

$$\log(freq(s,y)+1)$$

Die logarithmisierte absolute Frequenz zeigt an, auf wie vielen Vorkommen die berechnete relative Frequenz beruht und kann als Hinweis auf die Verlässlichkeit des berechneten Werts gelesen werden.

Sind für eine Sequenz S Kontextmerkmale basierend auf automatischen Annotationen berechnet worden und kommt diese Sequenz in den Trainingsdaten oder im umfangreichen Textkorpus direkt vor oder nach einer zu klassifizierenden Einheit vor, so wird das betreffende Kontextmerkmal erzeugt. Dieses Merkmal könnte beispielsweise darauf hinweisen, dass die zu klassifizierende Einheit von einem Kontext umgeben ist, der in den automatischen Annotationen bei jedem zweiten Vorkommen direkt vor einer bestimmten NE-Kategorie stand. Wird dieses Merkmal für eine Instanz des Trainingskorpus erzeugt, so gewichtet das anschließende SVM-Training dieses Merkmal. Wird das Merkmal bei einer erneuten Annotation des umfangreichen Korpus erzeugt, so führt es hoffentlich zu genaueren Klassifikationen.

Genau wie bei System I werden für alle Klassifikationsaufgaben dieselben Merkmale benutzt, d.h. es werden nicht nur die erweiterten Merkmale der aktuell zu erkennenden Klasse berücksichtigt. Damit hat der Klassifizierer für PERSON gewissermaßen auch Zugriff auf die erweiterten Merkmale für ORGANISATION, die beispielsweise aussagen, dass die zu klassifizierende Einheit von einem Kontext umgeben ist, der in den automatisch annotierten Daten sehr oft als Kontext von ORGANISATION vorkam. Dieses Merkmal kann auch für die Klassifikationsaufgabe PERSON hilfreich sein. Darüber hinaus ist es sehr viel effizienter, für mehrere Aufgaben auf dieselbe Instanzenrepräsentation zurückgreifen zu können. Zwar erfordert das Training weiterhin unterschiedliche Daten, doch unterscheiden sich diese nur hinsichtlich der Label, nicht aber hinsichtlich der Repräsentation der Instanzen.

Durch den Einsatz von individuellen Merkmalen für jeden Kontextausschnitt und jede NE-Klasse ist es möglich, dass ein Kontext Voraussagen für unterschiedliche NE-Klassen machen kann. Damit können mehrdeutige Kontexte angemessen repräsentiert werden. Darüber hinaus ist es möglich, dass für eine Wortfolge „a b c“ vor einem Fokus sowohl Werte für „c“, „b c“ und „a b c“ vorhanden sind. Diese getrennte Betrachtung eines rechten Kontextes wie etwa „ … lebt in Frankfurt“ bläht die Anzahl der Merkmale zwar auf, kann jedoch die aufsteigende Verlässlichkeit mit zunehmender Kontextlänge („lebt“, „lebt in“, „lebt in Frankfurt“) berücksichtigen und einen in seiner Gesamtheit noch nie erfassten rechten Kontext wie „lebt in Villingen-Schwenningen“ zumindest auf „lebt“ und „lebt in“ zurückführen.

Merkmale von Wörtern und Wortsequenzen zur Repräsentation von interner Evidenz

In System II werden die bereits in System I enthaltenen Merkmale zur wortweisen Repräsentation interner Evidenz erweitert, so dass es nun möglich ist, Merkmale über

Wortsequenzen zu verarbeiten. Wird durch den Basisklassifizierer etwa die Sequenz „Frauen helfen Frauen“ durch ein vorangestelltes „Verein“ als ORGANISATION erkannt, so wird sowohl für die vollständige Sequenz „Frauen helfen Frauen“ als auch die großgeschriebenen Bestandteile Merkmale für die betreffende Kategorie berechnet.

Die erweiterten Merkmale zur Repräsentation interner Evidenz basieren genau wie in System I zur Auswertung automatisch annotierter Daten auf diskretisierten Funktionswerten und logarithmisierten absoluten Frequenzen (vgl. 5.3.1). Diese Merkmale werden jedoch nicht nur für einzelne Wörter, sondern auch für ganze Sequenzen erzeugt.

Genau wie die Kontextmerkmale kann ein Wort oder eine Sequenz erweiterte Merkmale mehrerer NE-Kategorien besitzen, was die Berücksichtigung von Mehrdeutigkeiten ermöglicht. Ebenfalls wie bei den Kontextmerkmalen werden für alle Klassifikationsaufgaben dieselben Merkmale benutzt, so dass jeder Klassifizierer auch auf die erweiterten Merkmale der anderen Klassifizierer Zugriff hat.

Darüber hinaus wird für alle NEs, die aus mehr als einem Wort bestehen, ein zusätzlicher Eintrag über das erste und das letzte Wort der Sequenz abgelegt. Für alle Bestandteile r, die mindestens einmal als erste oder als letzte Einheit einer NE der Kategorie y vorkommen, wird ein Lexikoneintrag mit den folgenden zwei Werten erstellt:

$$\frac{freq(r,y)}{freq(r)} \quad \text{und} \quad \log(freq(r,y)+1)$$

Diese Merkmale dienen dazu, typische Beginn- und Endsequenzen mehrteiliger NEs zu erfassen: So sollten beispielsweise Vornamen ein hohes Gewicht für den Beginn mehrteiliger Personen-NEs erhalten; Wörter wie „Deutsche“, „Allgemeine“ etc. stehen oft am Beginn einer NE vom Typ ORGANISATION; „GmbH“ oder „AG“ sind oft am Ende einer NE vom Typ ORGANISATION.

Anpassung des Verfahrens: Eine Kombination von System I und System II

Während der Experimente mit System II stellte sich ein Befund heraus, der eine weitere Modifikation des Ansatzes erforderlich machte und der zu einer Kombination von System I und System II führte. Ein kleiner Vorgriff auf die Evaluation in Kapitel 6. ist für ein Verständnis der gewählten Kombination notwendig.

Die Experimente zeigten, dass die erweiterte N-Gram Modellierung im Vergleich zur klassischen N-Gram Modellierung eine höhere Precision bei leicht besserem F-Measure aufweist (Abschnitt 6.1.5) und sich somit theoretisch besser für die Ausnutzung nicht annotierter Daten eignet. Bei der anschließenden iterativen Annotation und Auswertung der umfangreichen nicht-annotierten Daten (Abschnitt 6.3.2) zeigt sich in der ersten Runde eine deutliche Verbesserung im Vergleich zum Basisklassifizierer, doch ist bei den weiteren Durchgängen nur noch eine geringe Steigerung zu beobachten (Abbildung 6.12). Allerdings zeigt eine Inspektion der SVM-Modelle eine interessante Beobachtung: Obwohl die Leistung der Modelle sich nur unerheblich verbessert, werden die Modelle bei jedem Durchgang kleiner bezüglich der Anzahl der Support Vektoren (Abbildung 6. 11). Zwar lässt sich die Anzahl von Support Vektoren nicht direkt interpretieren, aber als Indikator für die Modellgröße kann bei einer Reduzierung bei gleich bleibender bzw. leicht besserer Leistung von einer stärkeren Generalisierung ausgegangen werden. Die Vermutung liegt nahe, dass bei diesem Prozess spezifisches Wissen über einzelne NEs bzw. darin enthaltene Wörter und Wortsequenzen durch allgemeineres Wissen über Kontexte von NEs ersetzt wird.

Unter der Annahme, dass das System II, also mit den erweiterten N-Grammen und dem simultanen Erwerb von interner und externer Evidenz vorrangig zum Erwerb von Kontexten geeignet ist, aber System I, also der einzelwortbasierte Ansatz besser zum Erwerb interner Evidenz geeignet ist, wird folgende Kombination vorgeschlagen: Ausgehend von einem System II Basisklassifizierer wird solange simultan interne und externe Evidenz erworben, bis die Leistung sich nicht mehr verbessert und sich die Anzahl der Support Vektoren nicht mehr wesentlich weiter verringert. Darauf werden die extrahierten Kontextmerkmale für den weiteren Verlauf fixiert, also nicht mehr verändert. Die Merkmale zur internen Repräsentation hingegen werden so lange aktualisiert, bis der resultierende Klassifizierer keine Verbesserung mehr zeigt. Dabei werden genau wie in System I nur noch Einzelwort-Einträge genutzt.

5.3.3. Überblick über die erweiterten Merkmale und ihre Verwendung

Dadurch, dass die erweiterten Merkmale eines Worts bzw. einer Wortsequenz für alle NE-Kategorien sowohl interne als auch externe Evidenz bereitstellen können, kann pro Wort eine erhebliche Menge an Merkmalen zusammenkommen.

Für System I (vgl. 5.3.1) beschränken sich die Merkmale auf die interne Evidenz einzelner Wortformen. Pro Wortform und Kategorie ergeben sich maximal sieben Werte, nämlich die relative Frequenz der sechs diskretisierten $(\vec{w} \cdot \vec{x} + b)$-Werte und die logarithmisierte

absolute Frequenz (vgl. dazu die Abb. 5.8) des Vorkommens mit einem bestimmten NE-Label. Bei der Erkennung von drei NE-Klassen wie PERSON, ORGANISATION und LOCATION kann eine einzelne Wortform theoretisch ein Maximum von drei mal sieben Merkmalen tragen, auch wenn dies praktisch nur möglich wäre, wenn diese sehr oft mit allen drei Klassifikationen mit $(\vec{w} \cdot \vec{x} + b)$-Werten in allen gewählten Bereichen vorkäme. In System I werden die erweiterten Merkmale nicht nur für das Wort im Fokus, sondern für alle Wörter des Fensters erzeugt. Damit ergibt sich für ein Wortfenster der Größe sechs ausgehend von den maximal 21 Merkmalen pro Wort ein theoretisches Maximum von 126 erweiterten Merkmalen pro Instanz. Da erweiterte Merkmale jedoch nur für diejenigen Wortformen erzeugt werden, die mindestens einmal als NE vorkommen, werden für viele Wörter des Fensters gar keine erweiterten Merkmale eingesetzt.

In System II (5.3.2) kodieren die erweiterten Merkmale auch externe und nicht nur interne Evidenz, wobei letztere sich vom Ansatz in System I unterscheiden. System II hat zwar vielfältigere erweiterte Merkmale, das theoretische Maximum ist jedoch mit 69 Merkmalen geringer als bei System I. Die erweiterten Merkmale setzten sich wie folgt zusammen:

- Genau wie bei System I basieren die Merkmale der internen Evidenz auf den relativen Frequenzen der sechs diskretisierten $(\vec{w} \cdot \vec{x} + b)$-Werte und der logarithmisierten absoluten Frequenz (vgl. 5.3.1) des Vorkommens mit einem bestimmten NE-Label. Allerdings werden diese nicht nur für einzelne Wortformen, sondern auch für Wortsequenzen berechnet, aber nur noch für das Wort bzw. die Sequenz im Fokus erzeugt. Damit ergibt sich für diese Merkmale ein theoretisches Maximum von 21 Merkmalen pro Instanz. Dies ist deutlich weniger als in System I, in dem 21 Merkmale pro Fensterposition möglich sind.
- Darüber hinaus existieren auch jeweils zwei erweiterte Merkmale, welche für Wortformen einer Mehrwortsequenz erzeugt werden, wenn diese häufig am Beginn bzw. am Ende einer als NE klassifizierten Sequenz auftraten. Träfe dies sowohl für das Anfangs- als auch das Endwort der Sequenz für alle drei NE-Klassen zu, so ergäben sich weitere 12 Merkmale.
- Zur Abbildung der externen Evidenz kommen die Kontextmerkmale zum Einsatz. Jeweils zwei Merkmale liegen für alle Wortsequenzen der Länge eins bis drei vor, welche vor oder nach einem bestimmten NE-Label vorkamen und den in 5.3.2 beschriebenen Schwellwert erreichen (vgl. dazu die Abb. 5.10). Das Maximum an Kontextmerkmalen

läge vor, wenn alle drei Kontexte (d.h. der Kontext der Länge eins, zwei und drei) vor und alle drei Kontexte nach dem Fokus die zwei Einträge für alle drei Klassen hätten. Damit beträgt das Maximum 36 Kontextmerkmale.

5.4. Systemüberblick

Wie aus Abbildung 5.11 zu entnehmen ist, unterscheiden sich die erforderlichen Verarbeitungsschritte bei der Erzeugung und dem Einsatz der Klassifizierer nur in wenigen Punkten: Für die Erzeugung der erweiterten Merkmale bzw. Klassifizierer ist ein iteratives Durchlaufen der Schritte (4-6) zur Berechnung der erweiterten Merkmale erforderlich. Dies wurde mit System I (Abschnitt 5.3.1) und System II (Abschnitt 5.3.2) in zwei unterschiedlichen Varianten implementiert. Beide Varianten basieren auf der Ausgabe der SVM, also dem Wert von $(\vec{w} \cdot \vec{x} + b)$ und nicht auf den tatsächlichen NE-Annotationen. Diese beruhen auf dem eigentlichen Tagging (7) und dem Postprocessing (8-9), weshalb diese Schritte während der Erzeugung des Klassifizierers nicht benötigt werden. Ist die Berechnung der erweiterten Merkmale abgeschlossen, wird Schritt (4) nicht mehr benötigt und das System kann eingesetzt werden.

Verarbeitungsschritte	Erzeugen des NER-Modells	Einsatz des NER-Modells
(1) Vorverarbeitung (Tokenisierung, Satzgrenzen)	X	X
(2) Auswahl der Instanzen	X	X
(3) Einsetzen der Basismerkmale	X	X
(4) Iteration über 4-6 (siehe dazu Abbildung 5.8): Berechnung der erweiterten Merkmale aufgrund von automatischen Annotationen	X	-
(5) Einsetzen der erweiterten Merkmale	X	X
(6) SVM-Klassifizierung	X	X
(7) Tagging	-	X
(8) Postprocessing - Dokumenten-orientierte Klassifikation	-	X
(9) Weiteres Postprocessing	-	X

Abbildung 5.11: Systemüberblick – Verarbeitungsschritte beim Setup und Einsatz der Klassifizierer

Alle Verarbeitungsschritte außer der SVM-Klassifikation sind durch eines oder mehrere Perl-Skripte realisiert, welche aus den natürlichsprachlichen Texten Instanzen erzeugen und für

diese eine Vielzahl von Merkmalen extrahieren. Die Verarbeitungsschritte werden im Folgenden einzeln vorgestellt.

(1) Vorverarbeitung

Das Vorverarbeitungsskript überführt herkömmliche Textformate in ein tokenbasiertes Format. Dazu wird der Eingabetext mit Hilfe einer Kaskade von regulären Ausdrücken in einzelne Tokens zerlegt. Außerdem wird mit Hilfe bekannter Abkürzungen und Heuristiken unterschieden, ob es sich bei einem Punkt um das Satzende oder eine Abkürzung handelt.

(2) Auswahl der Instanzen

Aus der tokenbasierten Repräsentation sind diejenigen Wörter bzw. Wortsequenzen mitsamt Kontext auszuwählen, welche möglicherweise Teil einer NE sind (vgl. 5.2.1.). In Abhängigkeit von der Sprache und der zu erkennenden NEs kann hier eine Beschränkung auf gewisse Einheiten vorgenommen werden, wie etwa alle, die mit einem Großbuchstaben beginnen. Darüber hinaus sind hier die Heuristiken implementiert, welche die Anzahl der Instanzen beim Einsatz der erweiterten N-Gramme reduzieren (vgl. 5.2.1.).

(3) Einsetzen der Basismerkmale

Für die zu klassifizierende Einheit und den extrahierten Kontext werden die in 5.2.4 zusammengefassten Basismerkmale erzeugt. Das geschieht vorrangig mit regulären Ausdrücken. Die Verarbeitung mit der gewählten SVM Implementation ([SVM^{Light}]) erfordert die numerische Kodierung aller Merkmale. Die effiziente Umwandlung ist durchgehend als assoziative Liste implementiert. Für die Substring-Repräsentation (5.2.3) wird eine assoziative Liste aller Substrings des Trainingskorpus benutzt, welche die numerische Kodierung des Substringmerkmals zurückliefert.

(4) Berechnung der erweiterten Merkmale

Die Berechnung der erweiterten Merkmale ist nur bei der Erzeugung der Klassifizierer erforderlich. Dabei werden die Merkmale über mehrere Durchgänge der Verarbeitungsschritte 4-6 berechnet. Dieser Ansatz ist in zwei unterschiedlichen Versionen umgesetzt worden, welche in 5.3.1 und 5.3.2 ausführlich beschrieben sind. Für den Einsatz des Klassifizierers müssen die erweiterten Merkmale bereits zur Verfügung stehen. Allerdings spricht nichts dagegen, die Neuberechnung der erweiterten Merkmale periodisch zu wiederholen. Dies ist

dann sinnvoll, wenn sich das zu verarbeitende Korpus geändert hat, d.h. wenn wie bei einem Nachrichtenkorpus regelmäßig neue Texte hinzugefügt werden.

(5) Einsetzen der erweiterten Merkmale

Die erweiterten Merkmale haben jeweils spezifische Werte für jede Wortform (System I) oder auch für Wortsequenzen. Diese Merkmale können interne Evidenz wie etwa diskretisierte Ausgaben der SVM (ausführlicher in 5.3.1) oder externe Evidenz wie etwa die relative Frequenz des Vorkommens hinter oder vor bestimmten NE-Kategorien (5.3.2) bereitstellen. Die Merkmale der internen Evidenz werden für jedes Wort im Fenster erzeugt, die Kontextmerkmale hingegen nur, wenn sie tatsächlich als Kontext des aktuellen Fokus fungieren. Die erweiterten Merkmale werden mit insgesamt nur zwei assoziativen Listen verwaltet, welche jeweils für alle Kategorien Merkmale der internen bzw. externen Evidenz bereitstellen.

(6) SVM Klassifikation

Die gewählten Basismerkmale (5.2.4) und die erweiterten Merkmale (5.3.3) führen zu einem hochdimensionalen Raum, der allerdings sehr spärlich besetzt ist. Für die große Anzahl von Merkmalen ist die Zerlegung in positionale Substrings verantwortlich, die pro Position im Wortfenster zu jeweils mehreren tausend Merkmalen führt. Beim deutschen CoNLL-Korpus sind dies etwa 15’000 Substringmerkmale, was angewendet auf das Sechs-Wortfenster der traditionellen N-Gramme zu 90’000 potentiellen Merkmalen führt. In einer Instanz wird jedoch nur ein Bruchteil dieser Merkmale realisiert. Bei dem auf Basismerkmale (5.2.4) beschränkten Basisklassifizierer ergeben sich bei der traditionellen N-Gram-Modellierung maximal 66 Merkmale. Bei der erweiterten N-Gram Modellierung steigt diese Zahl um eins, da die Länge der Sequenz im Fokus als Merkmal kodiert ist. Für jedes zusätzliche Wort im Fokus kommen 11 Basismerkmale dazu, vorausgesetzt, dass das zusätzliche Wort im Fokus keine gemeinsamen Merkmale mit den anderen Wörtern im Fokus hat. Befindet sich also eine Sequenz von vier Wörtern im Fokus eines erweiterten N-Grams, so ergäbe sich ein Maximum von 99 Merkmalen. Die erweiterten Klassifizierer arbeiten auf demselben Merkmalsraum wie die Basisklassifizierer, zuzüglich der erweiterten Merkmale (Abschnitt 5.3.3). Dies sind bei System I 126, bei System II hingegen 69 zusätzliche Merkmale.

Der Umgang mit diesem hochdimensionalen, aber spärlich besetzten Merkmalsvektor erfordert einen äußerst leistungsfähigen Lernalgorithmus. Außerdem hat der Lernalgorithmus

eine gewisse Geschwindigkeitsanforderung zu erfüllen. Dieses ist zum einen aus anwendungspraktischer Sicht notwendig, zum anderen erfordert die geplante mehrfache automatische Annotierung eines umfangreichen Korpus einen Klassifizierer, der für die Verarbeitung mehrerer Millionen Wörter möglicherweise mehrere Stunden, aber nicht mehrere Tage benötigt. Die Wahl fiel auf die lineare Support Vektor Maschine (SVM), da sich diese durch einen effizienten Umgang mit hochdimensionalen, aber dünn besetzten Merkmalsvektoren auszeichnet. Die SVM ist bereits in 4.2.5 ausführlich besprochen worden, weswegen hier nur noch der praktische Einsatz vorzustellen ist.

Wie in 4.2.5 ausgeführt, löst eine SVM immer nur eine binäre Klassifikationsaufgabe. NER ist jedoch ein Mehrklassenproblem, so dass mehrere SVMs und einige Überlegungen zur Kombination derselben erforderlich sind.

Im Rahmen des Korpus-adaptiven Ansatzes wurde die Aufgabenstellung folgendermaßen gewählt:

Für eine Folge von Beobachtungen,

$X = x_1,\ldots,x_l$ in unserem Fall Wortvorkommen in einem Text[2], die möglicherweise eine NE sind, ist die entsprechende Folge von NE-Labels

$Y = y_1,\ldots,y_m$ gesucht, wobei im Falle des CoNLL-Korpus $y_i \in$ {PER, ORG, LOC, O} gilt.

Jedes x wird durch einen Merkmalsvektor $\vec{x}$ repräsentiert. Die Erzeugung des Merkmalsvektors wird durch eine Menge von Merkmalsfunktionen F geleistet:

$F : x \mapsto \vec{x}$, wobei $\vec{x}_i = f_i(x)$ und x_i die i-te Komponente von $\vec{x}$ ist und das Ergebnis der Merkmalsfunktion f_i ist.

Die Merkmalsfunktion erzeugt sowohl die Basismerkmale (5.2.4) als auch die erweiterten Merkmale (5.3.3).

Wie bereits in Kapitel 4.2.5 ausgeführt, finden sich zwar viele Überlegungen für die effiziente Anwendung der SVM auf Multiklassenprobleme (vgl. dazu [HSU & LIN 2001]), doch bleiben diese weitestgehend ohne Einfluss auf den SVM-Einsatz für die NER. Auch in eigenen Vorversuchen hat sich immer wieder gezeigt, dass die einfache „ein-Klassifizierer-pro-NE-

[2] An allen anderen Stellen bezeichnet x_i die i-te Komponente des Merkmalsvektors $\vec{x}_i$.

Klasse“-Strategie, in der nur die Instanzen der jeweiligen Klasse als positive Instanzen betrachtet werden, zu den besten Ergebnissen führt. Ausgehend von der Anzahl m der Label Y, wird eine Menge von $m-1$ SVMs K benötigt. Im Falle des CoNLL-Korpus drei SVMs:

$$k_{PER}(\vec{x}) = \vec{w}_{PER} \cdot \vec{x} + b_{PER}$$

$$k_{ORG}(\vec{x}) = \vec{w}_{ORG} \cdot \vec{x} + b_{ORG}$$

$$k_{LOC}(\vec{x}) = \vec{w}_{LOC} \cdot \vec{x} + b_{LOC}$$

Der jeweilige Wert von $k(\vec{x})$ bildet dabei die Grundlage für den siebten Verarbeitungsschritt, das eigentliche Tagging.

(7) Tagging

Die Ausgaben der $m-1$ SVMs können nicht einfach mittels $sign(\vec{w} \cdot \vec{x} + b)$ aufgelöst werden. Die Binarisierung des Multiklassenproblems NER kann dazu führen, dass mehrere SVMs einen positiven Wert von $(\vec{w} \cdot \vec{x} + b)$ berechnen. Dies muss aufgelöst werden, damit ein Wortvorkommen nicht mehreren NE-Klassen zugehörig ist. Die eigentliche Tagging-Funktion u löst solche Schwierigkeiten nach dem Maximum der oben eingeführten K auf, welche dem jeweiligen Wert von $(\vec{w} \cdot \vec{x} + b)$ entsprechen:

$$u(\vec{x}) = \left\{ \begin{array}{l} y_i \\ \text{falls } k_{yi}(\vec{x}) > 0 \wedge k_{yi}(\vec{x}) = \max\limits_{j=1,..,m-1} k_{yi}(\vec{x}) \\ 0 \text{ sonst} \end{array} \right\}$$

Im Falle der erweiterten N-Gram Modellierung kommt es vor, dass ein Wort mehrmals als Bestandteil unterschiedlicher Sequenzen klassifiziert wird. Hierbei kann es etwa zu folgenden, sich widersprechenden Voraussagen kommen:

w-3	w-2	w-1	Fokus	w+1	w+2	Klasse
...	...	die	Sparkasse	Hamburg	...	**O**
...	...	die	Sparkasse Hamburg	wird	...	**ORG**
...	die	Sparkasse	Hamburg	wird	...	**LOC**

Daher wird bei der erweiterten N-Gram Modellierung zuerst ein „Longest-Match-Kritierium“ benutzt: Berechnen eine oder mehrere SVMs für eine bestimmte Position im Text mehrere Sequenzen mit einem Wert von $\vec{w} \cdot \vec{x} + b > 0$, so wird diejenige Sequenz ausgewählt, welche eine größere Anzahl von Tokens abdeckt. Führt dies nicht zu einem eindeutigen Ergebnis, so wird die längste Sequenz mit dem höchsten $\vec{w} \cdot \vec{x} + b$ ausgewählt.

(8) Postprocessing - Dokumenten-orientierte Klassifikation

Nachdem die Ausgaben der verschiedenen SVMs zum Tagging des Textes benutzt worden sind, kann das Ergebnis der NER durch weitere Postprocessing-Komponenten optimiert werden. Die wichtigste und am weitesten verbreitete dieser Nachbearbeitungen ist die Dokumenten-orientierte Klassifikation. Für die Dokumenten-orientierte Klassifikation wird das Phänomen ausgenutzt, dass Wortformen innerhalb einer zusammengehörigen Texteinheit meist nur in einer Verwendung benutzt werden. Das Phänomen wurde in [GALE ET AL. 1992] im Zusammenhang mit der Disambiguierung semantischer Lesarten festgestellt. Für die NER werden dazu die prädiktiven NE-Vorkommen benutzt, um nicht-prädiktive Vorkommen zu klassifizieren (vgl. dazu Abschnitt 3.2.6).

Das Verfahren wird wie in den meisten NER-Systemen in Form eines dynamisch erzeugten temporären Lexikons eingesetzt. Dazu werden in einem ersten Schritt alle innerhalb einer Texteinheit automatisch erkannten NEs in einem dynamischen Lexikon gespeichert. Danach werden alle weiteren Vorkommen der im temporären Lexikon gespeicherten Wörter mit der im Lexikon eingetragenen NE-Klasse markiert. Wortformen, die innerhalb einer Texteinheit mit unterschiedlichen NE-Klassen ausgezeichnet wurden, werden nicht in das dynamische Lexikon aufgenommen.

In Abhängigkeit der Sprache bzw. Domäne des zu verarbeitenden Korpus kann es sinnvoll sein, das dynamische Lexikon auf großgeschriebene Einheiten zu beschränken und/oder gewisse häufig vorkommende Wortformen von der Aufnahme grundsätzlich auszuschließen. Für die Experimente mit dem deutschen CoNLL-Korpus etwa wurden kleingeschriebene Wörter nur als Teil einer Mehrworteinheit ins dynamische Lexikon aufgenommen, während alle großgeschriebenen Wörter sowohl als Mehrwort als auch als Einzelworteintrag gespeichert wurden.

Mit der Strategie der Dokumenten-orientierten Klassifikation werden also keine neuen, bisher unbekannten NEs entdeckt, vielmehr werden weitere Vorkommen von gefundenen Namen markiert. Das Verfahren beachtet den Kontext eines Wortes nicht, sondern geht davon aus,

dass eine Wortform innerhalb einer Texteinheit meist in derselben Verwendung benutzt wird. Deshalb wird eine gewisse Fehlerquote in Kauf genommen, so dass zu erwarten ist, dass das Verfahren positiv auf den Recall, möglicherweise aber leicht negativ auf die Precision Einfluss hat.

Die Strategie der Dokumenten-orientierten Klassifikation ist davon abhängig, dass dem NER-System die Texteinheiten bzw. Segmente zur Verfügung stehen. Für die in Kapitel 6. durchgeführten Experimente bzw. die dafür bearbeiteten Korpora waren diese Informationen zugänglich. Im CoNLL-Korpus wurde als Dokument bzw. Texteinheit jeweils ein Zeitungsartikel festgelegt, im Falle des biomedizinischen Korpus ([JNLPBA-2004]) jeweils nur ein Medline-Abstract ([MEDLINE]), also die Zusammenfassung eines Fachartikels. Sind in einem Korpus keine Texteinheiten erkennbar, so wäre es immer noch überprüfenswert, ob das Verfahren nicht auch mit festen Wortfenstern beispielsweise der Größe 100 Wörter davor und 500 Wörter danach eingesetzt werden kann. Dieses Vorgehen wurde von [VOLK & CLEMATIDE 2001] vorgeschlagen.

(9) Weiteres Postprocessing

In derselben Verarbeitungskomponente wie die Dokumenten-orientierte Klassifikation (8) werden auch Koordinationsphänomene ausgenutzt. Dabei werden Folgen koordinierter NEs, also eine längere Aufzählung von NEs ermittelt. Für das deutsche CoNLL-Korpus wurde unter folgenden Kriterien eine Koordination der Teilsequenzen angenommen:

- Wenn der Satz mindestens vier aus maximal zwei großgeschriebenen Wörtern bestehende Teilsequenzen enthält, die alle durch dasselbe Satzzeichen miteinander verbunden sind, wobei die letzte Sequenz auch durch ein „und“ angebunden sein darf
- und mindestens drei Teilsequenzen als NEs derselben Kategorie ausgezeichnet sind.

Diese Regel entdeckt etwa die aufgezählten Mitglieder einer Fußballmannschaft, aber auch Aufzählungen von Ortsnamen. Die aufgezeigte Ausnutzung von Koordinationsphänomenen ist auf jeden Fall Einzelsprachen-, möglicherweise sogar Korpus-spezifisch. Allerdings ist es denkbar, während der Anpassung eines NER-Systems gezielt nach Koordinationsstrukturen zu fahnden und diese anschließend zu berücksichtigen. In Abhängigkeit von Eigenschaften der zu verarbeitenden Texte kann die optimale Ausnutzung von Koordinationsphänomenen einen zentralen Beitrag zur NER darstellen. Koordinierte Einheiten haben als Kontext weitere koordinierte Einheiten, nicht aber einen Kontext, also externe Evidenz, die durch N-

Gramme gelernt werden kann (vgl. dazu 3.2.3). Enthält ein Text also umfangreiche Aufzählungslisten mit NEs, so können die Einträge dieser Listen nur durch interne Evidenz oder Dokumenten-orientierte Klassifikation korrekt klassifiziert werden. Liegt jedoch nicht genügend interne Evidenz vor und kommen diese Einheiten an keiner anderen Stelle im Dokument vor, so werden diese ohne die Ausnutzung des Koordinationsphänomens nicht korrekt behandelt.

Der Verarbeitungsschritt (9) ist der bevorzugte Ort, um weitere Postprocessing-Komponenten zu integrieren, wie beispielsweise die in [RÖSSLER 2004a] beschriebene. Um die Adaptivität des hier vorgeschlagenen Systems zu prüfen, wurde ein System zur biomedizinischen NER entwickelt (vgl. 6.2). Bei der Analyse der ersten Resultate zeigte sich, dass vor allem die Länge der NEs und die darin enthaltenen Stopp-Wörter die Erkennungsrate drastisch senkten. Um dies abzufangen, wurde eine Postprocessing-Komponente aufgesetzt, die sowohl auf den Tags als auch auf den $(\vec{w} \cdot \vec{x} + b)$ - Werten der SVMs operiert. Auch wenn die Komponente die Ergebnisse leicht verbesserte, so ist die Gestaltung der Komponente doch sehr pragmatisch und wenig geeignet, die Grundlage einer eleganten und allgemeingültigen Lösung des Problems zu werden. Das durch die Komponente angegangene Problem langer NEs, die darüber hinaus Stopp-Wörter enthalten, ist vermutlich angemessener durch Sequenz-orientierte Verarbeitungsstrategien wie etwa CRFs (4.2.4) oder SVM-HMM (4.2.5) zu lösen. Nichtsdestotrotz kann es in einem NER-System notwendig sein, fehlende allgemeingültige Lösungsstrategien durch pragmatische Komponenten zu ersetzen, die meist als Postprocessing-Komponenten in Verarbeitungsschritt (9) realisiert werden.

6. Evaluation

Um das in Kapitel 5 vorgestellte System zu evaluieren, wird eine Vielzahl von Experimenten durchgeführt. Dazu wird das System in unterschiedlichen Konfigurationen trainiert und anschließend evaluiert. Die Implementation ist durch eine Reihe von Perl-Skripten realisiert, die gut auf die in 5.4 beschriebenen Verarbeitungsschritte abgebildet werden können. Sowohl zum Lernen als auch zur Klassifikation wird die Implementation SVM^{Light} in der Version 5.0 [SVM^{Light}] eingesetzt. Die Ausgabe der SVM-Klassifizierer wird wiederum von Perl-Skripten als Annotation auf den Ausgangstext gelegt.

Der Kern des Korpus-adaptiven Systems ist der Einsatz nicht annotierter Daten. Er ist von der Idee motiviert, dass ein einfacher Klassifizierer anhand der Auswertung seiner eigenen Klassifikationsentscheide verbessert werden kann (5.3). Dazu wird der sog. Basisklassifizierer, der aufgrund von manuell annotierten Daten trainiert worden ist, zur Klassifikation eines umfangreichen nicht-annotierten Korpus eingesetzt. Basierend auf den Klassifikationsentscheiden werden Merkmale berechnet, welche interne oder externe Evidenz zur NER ableiten. Mithilfe dieser sog. erweiterten Merkmale wird der Basisklassifizierer zum erweiterten Klassifizierer. Zeigt dieser erweiterte Klassifizierer eine verbesserte Erkennungsrate, so wird er erneut zur Klassifikation des umfangreichen nicht-annotierten Korpus eingesetzt und die erweiterten Merkmale werden erneut berechnet. Dieser Vorgang wird solange wiederholt, bis keine Verbesserung mehr zu erkennen ist.

Die unterschiedlichsten Aspekte des Basisklassifizierers werden in Abschnitt 6.1 evaluiert. Die beabsichtigte Adaptivität des Verfahrens wird in 6.2 anhand der Erkennung biomedizinischer NEs in englischen Fachtexten untersucht. Die Wirkung der erweiterten Merkmale, das Zusammenspiel von möglichen und eingesetzten weiteren Komponenten mit dem Gesamtsystem und das Laufzeitverhalten wird in 6.3 evaluiert. Eine abschließende Diskussion der Ergebnisse findet in 6.4 statt.

Wenn nicht anders vermerkt, wurden die hier beschriebenen Experimente auf dem deutschen CoNLL-Korpus durchgeführt. Die Annotationen des Korpus wurden in Abschnitt 2.2.2 diskutiert. Angaben zum verwendeten Trainings- und Testkorpus finden sich in Abbildung 6.1. Die Kategorie MISC (MISCELLANEOUS), welche eine Restgruppe für Namen darstellt, die nicht zu den Kategorien PERSON (PER), ORGANISATION (ORG) oder LOCATION (LOC) gehören, wurde nur für einen Teil der Experimente berücksichtigt. Da die Grenzen dieser Kategorie anhand der manuellen Annotationen inkonsistent und nicht nachvollziehbar

erschienen, wurde vollständig darauf verzichtet, Evidenz für diese Kategorie aus nicht-annotierten Daten abzuleiten und alle Annotationen der besagten Kategorie wurden für alle Experimente mit dem erweiterten Klassifizierer sowohl aus Trainings- als auch Testdaten entfernt. Bei der Evaluation des Basisklassifizierers hingegen wurde auch die Kategorie MISC berücksichtigt.

	PER	ORG	LOC	MISC	Wörter
Trainingskorpus (CoNLL-Datei deu.train)	2773	2426	4360	2269	220189
Testkorpus (CoNLL-Datei deu.testa)	1400	1241	1179	995	54173

Abbildung 6.1: Anzahl NEs und Wörter in den manuell annotierten Daten

Alle Experimente auf CoNLL-Daten wurden mit dem Skript „conlleval“ evaluiert, welches zusammen mit den CoNLL-Daten erhältlich ist. Die Evaluation benutzt ein exaktes Treffer-Maß, welches nur vollständige Treffer als korrekt berücksichtigt. Das für die Experimente auf den biomedizinischen Daten eingesetzte Skript benutzt ebenfalls ein exaktes Treffer-Maß und wird zusammen mit den betreffenden Daten ausgeliefert. Um die Ergebnistabellen übersichtlich zu gestalten, wurden in den meisten Fällen nur die Angaben zum F-Measure abgebildet. Die vollständigen Ergebnisse sind im Anhang enthalten.

6.1. Basisklassifizierer

In diesen Experimenten werden Bestandteile des Basisklassifizierers evaluiert. Der Basisklassifizierer stellt den Einstieg in das in dieser Arbeit vorgestellte Verfahren zur Korpus-adaptiven NER dar. Die Basismerkmale, also die Merkmale auf denen der Basisklassifizierer operiert, sind in Kapitel 5.2 beschrieben und in 5.2.4 zusammengefasst. Von den im Rahmen des Systemüberblicks in 5.4 aufgeführten Verarbeitungsschritten werden im Basisklassifizierer nur die Schritte (1-3) und (6-7) durchgeführt, d.h. es fehlen die erweiterten Merkmale und abgesehen vom Tagging findet nach der SVM-Klassifikation kein weiteres Postprocessing mehr statt.

Für diese Experimente wird von der Konfiguration ausgegangen, wie sie in [RÖSSLER 2004b] dargestellt ist und als Grundlage für die in [RÖSSLER 2004a] und [RÖSSLER & MORIK 2005] beschriebenen Ansätze dient.

In allen Experimenten wird dieselbe Konfiguration des Klassifizierers verwendet, abgesehen von den zu evaluierenden Aspekten. Diesen Aspekten werden alternative Konfigurationen gegenübergestellt und experimentell verglichen. Dazu sollen anhand von Experimenten folgende Fragen untersucht werden:

- **Substring-Repräsentation:** Wie ist die gewählte Repräsentation mit positionalen Substrings der Repräsentation der gesamten Wortform und alternativen Substring-Repräsentationen gegenüber zu bewerten?
- **Größe des betrachteten Kontexts:** Lässt sich die Größe des betrachteten Ausschnittes, also drei Einheiten vor und zwei nach der zu klassifizierenden Einheit experimentell rechtfertigen?
- **Lineare SVMs:** Wie wirkt sich der Verzicht auf die oft mächtigeren, nicht-linearen Trennfunktionen in den Experimenten aus?
- **Wortoberflächenmerkmale und Wortlänge:** Was ist der Einfluss solcher datenorientierten Generalisierungen auf die Erkennungsrate?
- **Erweiterte N-Gram Modellierung:** Lässt sich die vorgeschlagene sequenzorientierte Erweiterung der N-Gram Modellierung effizient implementieren, und wie ist diese Erweiterung im Vergleich zu den klassischen N-Grammen zu beurteilen?

6.1.1. Evaluation der gewählten Substring-Repräsentation

Um die gewählte Substring-Repräsentation (5.2.3) zu evaluieren, wird diese anderen Substring-Zerlegungen und der Repräsentation der kompletten Wortformen gegenübergestellt. Dazu werden anhand des gewählten Trainingskorpus die Merkmale erzeugt und anschließend evaluiert. Zum Vergleich wurden Experimente mit den folgenden Konfigurationen durchgeführt.

(1) **Substrings, positional:** Dies entspricht dem in 5.2.3 beschriebenen Vorgehen, welches in allen anderen Experimenten eingesetzt wurde. Dazu werden von den Wörtern des Trainingskorpus das letzte Unigram, das letzte Bigramme und die ersten und die letzten drei Trigramme zusammen mit der Positionsangabe als Merkmale definiert. Auf dem CoNLL-Korpus führt dies zu 14218 Merkmalen.

(2) **Wortform:** Hierbei werden die Wortformen nicht zerlegt, sondern jede Wortform stellt ein Merkmal dar. Auf dem Trainingskorpus führt dies zu 32930 Merkmalen, also mehr als doppelt so viele wie im gewählten Ansatz mit den positionalen Trigrammen. Durch die

Einschränkung eines Frequenzschwellwerts, dass nur diejenigen Wortformen zu einem Merkmal führen, die im Trainingskorpus mehr als einmal vorkommen, wurde eine vergleichbare Anzahl von 12529 Merkmalen erzeugt. Die Merkmalsmenge wurde sowohl mit (2a) als auch ohne (2b) Frequenzschwellwert evaluiert.

(3) **Bigramm, positional:** Anstelle von positionalen Trigrammen bilden Bigramme in analoger Art und Weise die Merkmale zur Repräsentation der Wörter. Auf dem gewählten Trainingskorpus führt dies zu 4128 Merkmalen. Die sehr viel geringere Anzahl von Merkmalen gegenüber den positionalen Trigrammen ist damit zu erklären, dass aus der gegebenen Zeichenmenge, nämlich den Buchstaben, Ziffern, Sonderzeichen etc. sehr viel mehr Dreierkombinationen als Zweierkombinationen möglich sind.

(4) **Trigramm, nicht positional:** Hierbei bilden alle möglichen Trigramme aus einem Wort Merkmale. Ob ein Substring ein- oder mehrmals vorkommt, wird nicht berücksichtigt. Zusätzlich wird das erste, das letzte und die zwei letzten Zeichen als Merkmal repräsentiert. Auf dem gewählten Trainingskorpus führt dies zu 11024 Merkmalen.

(5) **Bi- und Trigram, nicht positional:** Genau wie (4), nur dass hier zusätzlich alle möglichen Bigramme mitbetrachtet werden, was zu 12455 Merkmalen führt.

(6) **Bigram, nicht positional:** Genau wie (4), nur dass hier anstelle der Trigramme nur alle möglichen Bigramme betrachtet werden, was zu 2153 Merkmalen führt.

(7) **Keine Merkmale für den String:** Um einen Eindruck vom Einfluss der Repräsentation der Zeichenketten zu gewinnen, wurde das Modell für das letzte Experiment nur mit den Merkmalen der Wortoberfläche und der Wortlänge trainiert. Eine solche Repräsentation stellt äußerst wenig Evidenz zur NER bereit, gibt aber Hinweise zum Einfluss der Wortoberflächenmerkmale.

(8) **Keine Merkmale für den String – Englisch:** Um die Evaluation der Wortoberflächenmerkmale besser beurteilen zu können, wurde dasselbe Experiment mit den englischen Texten der CoNLL-Daten durchgeführt.

	Merkmale pro Fenster-position	LOC F-Measure	PER F-Measure	ORG F-Measure	MISC F-Measure	Gesamt Precision	Gesamt Recall	Gesamt F-Measure
(1) Substrings, positional (gewähltes Verfahren)	14218	50.60	42.67	**45.38**	**17.98**	**70.22%**	29.28%	41.33
(2a) Wortform, Frequenz>1	12529	51.91	26.45	43.39	2.72	68.72%	22.86%	34.31
(2b) Wortform, Frequenz>0	32930	**52.36**	27.32	43.60	2.72	69.13%	23.22%	34.76
(3) Bigramm, positional	4128	48.25	40.40	43.16	10.24	67.91%	26.71%	38.34
(4) Trigramm, nicht positional	11024	51.61	**46.65**	45.33	14.70	69.54%	**30.52%**	**42.42**
(5) Bi- und Trigram, nicht positional	12455	51.50	45.29	44.90	13.89	68.99%	29.92%	41.74
(6) Bigramm, nicht positional	2153	46.98	40.21	42.44	4.05	65.58%	25.74%	36.97
(7) keine Repräsentation der Wortform (Deutsch)	0	2.00	0.00	38.01	0.00	61.43%	7.28%	13.02
(8) keine Repräsentation der Wortform (Englisch)	0	44.44	50.37	42.01	0.22	67.33%	29.62%	41.14

Abbildung 6.2: Ergebnisse mit unterschiedlichen Wort-Repräsentationen; das jeweils beste Ergebnis ist fett gedruckt.

Abbildung 6.2 zeigt die Ergebnisse der durchgeführten Experimente. Zur Diskussion sind neben der eigentlichen Erkennungsleistung auch die Anzahl der resultierenden Merkmale und der erforderliche Verarbeitungsaufwand zu berücksichtigen. Die Anzahl der Merkmale pro Fensterposition ist in der zweiten Spalte angegeben. Die für das Verfahren gewählte Größe von sechs Wörtern im Kontextfenster ergibt also für Ansatz (2b) annähernd 200.000 potentielle Merkmale zur Repräsentation der Instanzen.

Wird das gewählte Verfahren (1), die positionalen Substrings, der Repräsentation der Wortformen (2a-2b) gegenübergestellt, so ist die wortbasierte Merkmalserzeugung einzig bei der Kategorie Ortsangaben leicht besser. Der große Unterschied beim F-Measure über alle Kategorien zeigt jedoch klar die Überlegenheit des gewählten Verfahrens, selbst wenn wie in (2b) ein äußert großer Merkmalsraum gewählt wird. Offensichtlich eignet sich die gewählte Zerlegung in Substrings besser zur Generalisierung über den Lerninstanzen, auch wenn sich dieser Effekt bei den NE-Klassen unterschiedlich zeigt. Beeindruckend ist hierbei vor allem der große Vorsprung bei der Kategorie PERSON.

Auch der Vergleich mit den positionalen Bigrammen in (3) fällt eindeutig zu Gunsten von (1) aus. Allerdings führt (3) zu sehr viel weniger Merkmalen. Leicht unterlegen zeigt sich (1) hingegen den nicht-positionalen Trigrammen in (4). Diese sind insbesondere im Gesamt-F-Measure leicht besser als bei (1). (1) ist bei der Gesamtpräzision und bei ORGANISATION in ganz geringem Maße besser als (4). Etwas deutlicher, wenn auch auf sehr tiefem Niveau, fällt der Unterschied bei der Kategorie MISCELLANEOUS aus. Allerdings kann für den Einsatz von (1) trotz leicht schwächeren Ergebnissen mit Laufzeiten argumentiert werden. Die positionalen N-Gramme beschränken sich auf die ersten und die letzten drei Positionen im String, während die nicht-positionalen N-Gramme in der hier benutzten Form für alle Positionen extrahiert werden. In Untersuchungen zur Laufzeit zeigt sich denn auch, dass der Einsatz von (4) etwa 50% mehr Rechenzeit benötigt als (1).

(5) ist vor allem im Vergleich zu (4) interessant, deutet es doch an, dass der zusätzliche Einsatz von Bigrammen die Evidenz von Trigrammen nicht unterstützt, sondern sogar zu leicht schlechteren Ergebnissen führt.

Das Experiment in (7) macht vor allem deutlich, dass NER für deutsche Texte ohne spezifisches Wissen über Wörter nahezu unmöglich ist. Dies zeigt auch der Vergleich mit den englischen Daten in (8). Hierbei genügen offensichtlich Wortlänge und Wortoberflächenmerkmale, um eine Identifikationsleistung zu erhalten, die derjenigen des Basisklassifizierers in deutschen Texten entspricht. Äußerst erstaunlich in (7) ist einzig die Kategorie ORGANISATION, die im Vergleich zu (1)-(6) keinesfalls dramatisch abfällt. Eine kleine Inspektion der Daten deutet darauf hin, dass dies vorrangig an Organisationsnamen liegt, die einzig aus Großbuchstaben zusammengesetzt sind. Mit diesem durch die Wortoberflächenmerkmale (vgl. Abschnitt 5.2.2) abgedeckten Kriterium können viele Organisationsnamen entdeckt werden, wenn dabei auch eine hohe Fehlerrate in Kauf genommen werden muss. Allerdings bedeuten die Ergebnisse in (7) auch, dass die Substring-Merkmale des gewählten Basisklassifizierers (1) die Erkennungsleistung für die Kategorie ORGANISATION nur in geringem Maße verbessern.

6.1.2. Evaluation des gewählten N-Gram Kontexts

Um den Kontext der zu klassifizierenden Einheiten zu berücksichtigen, wurde in Abschnitt 5.2.1 eine traditionelle N-Gram Modellierung vorgeschlagen, welche die drei vorhergehenden und die zwei nachfolgenden Wörter bzw. Tokens berücksichtigt. Um den gewählten

Kontextausschnitt zu evaluieren, wurden folgende Kontext-Konfigurationen miteinander verglichen:

- Drei Einheiten davor, zwei dahinter: Dies entspricht dem in Abschnitt 5.2.1 gewählten Vorgehen.
- Zwei Einheiten davor, zwei dahinter: Die Konfiguration macht den Merkmalsraum ein wenig kleiner, da eine Einheit weniger betrachtet wird.
- Drei Einheiten davor, drei dahinter: Die Konfiguration betrachtet auf der rechten Seite einen größeren Kontext und erlaubt deshalb, mehr externe Evidenz zu berücksichtigen.
- Kein berücksichtigter Kontext: Um den Einfluss des Kontexts zu evaluieren, wird nur das zu klassifizierende Wort ohne weiteren Kontext betrachtet.

	LOC F-Measure	PER F-Measure	ORG F-Measure	MISC F-Measure	Gesamt Precision	Gesamt Recall	Gesamt F-Measure
Kontext-Window drei davor, zwei dahinter (gewähltes Verfahren)	**50.60**	42.67	45.38	17.98	**70.22%**	**29.28%**	**41.33**
Kontext-Window zwei davor, zwei dahinter	49.23	**42.85**	**46.07**	17.80	69.92%	29.20%	41.19
Kontext-Window drei davor, drei dahinter	49.10	42.07	45.63	16.05	69.80%	28.55%	40.53
kein Kontext-Window	49.01	27.93	23.09	**28.69**	47.23%	25.20%	32.87

Abbildung 6.3: Ergebnisse mit unterschiedlichen Kontextausschnitten; das jeweils beste Ergebnis ist fett gedruckt.

Die Evaluationsergebnisse in Abbildung 6.3 zeigen, dass das gewählte Kontextfenster eine bessere Erkennungsleistung bringt, wenn auch mit äußerst geringen Unterschieden. In den nur flüchtig dokumentierten vorbereitenden Experimenten war dieser Unterschied sehr viel deutlicher. Offensichtlich wurde dieser durch Änderungen in anderen Bereichen, möglicherweise durch eine Überarbeitung der Wortoberflächenmerkmale, abgeschwächt. Die Ergebnisse sind ein deutlicher Hinweis darauf, dass die isolierte Evaluation von Bestandteilen zur Fixierung von Parametern zwar unumgänglich ist, doch dabei Wechselwirkungen mit anderen Bestandteilen leicht übersehen werden können.

Der Einfluss des Kontexts kann an den Ergebnissen des letzten Experiments abgelesen werden. Der Recall über alle Werte sinkt zwar ab, doch ist dieser Einfluss am stärksten bei der Precision messbar. Das geht mit der Annahme einher, dass der Kontext zwar auch zur Entdeckung von NEs, vor allem aber zur Disambiguierung notwendig ist. Dies gilt auch für die Kategorie LOCATION, deren F-Measure zwar kaum einen Unterschied zeigt, deren Precision jedoch von 78% bei der gewählten Kontextberücksichtigung auf 56% gefallen ist (siehe dazu die detaillierten Ergebnisse im Anhang). Auffällig ist, dass bei der Kategorie MISCELLANEOUS ohne Kontext die mit Abstand besten Ergebnisse erzielt worden sind. Dies ist möglicherweise damit zu erklären, dass diese „Restgruppe“ nicht in ähnlichen Kontexten vorkommt, da es eine äußerst heterogene Klasse ist.

6.1.3. Die Evaluation des SVM-Kernels

Wie in Abschnitt 5.4 beschrieben, basiert das Verfahren auf linearen SVMs. Der Vorteil linearer gegenüber nicht-linearen Kernfunktionen liegt in der deutlich höheren Verarbeitungsgeschwindigkeit. Allerdings sind nicht alle Klassifikationsaufgaben linear separabel, so dass nicht-lineare SVMs möglicherweise zu sehr viel besseren Klassifikationsleistungen führen. Bei äußerst hochdimensionalen Vektorräumen, wie den hier zu verarbeitenden, ist die richtige Wahl des Kernels ohne experimentelle Befunde kaum möglich. In Experimenten wurde der gewählte lineare Kernel mit dem polynomialen Kernel verglichen.

	LOC F-Measure	PER F-Measure	ORG F-Measure	MISC F-Measure	Gesamt Precision	Gesamt Recall	Gesamt F-Measure
Linearer Kernel (gewähltes Verfahren)	50.60	42.67	45.38	17.98	70.22%	29.28%	41.33
Polynomialer Kernel	42.31	36.35	41.07	9.16	73.49%	22.59%	34.56

Abbildung 6.4: Die Ergebnisse des linearen und des polynomialen Kernels im Vergleich

Abbildung 6.4 zeigt die Ergebnisse der Evaluation. Interessanterweise scheinen die Daten sehr viel besser mit einem linearen Kernel trennbar zu sein. Es zeigt sich ein durchgehend höheres F-Measure über alle Kategorien. Die leicht höhere Precision des polynomialen

Kernels ist vor dem Hintergrund des deutlich schwächeren Recalls unerheblich. Das Ergebnis ist insbesondere zufrieden stellend, da der Zeitaufwand für die Experimente mit dem polynomialen Kernel noch einmal deutlich gemacht hat, dass ein Einsatz nicht-linearer Kernelfunktionen aufgrund des erforderlichen Zeitaufwandes absolut undenkbar ist. Die für den Ansatz essentielle Annotation umfangreicher Korpora (vgl. Abschnitt 5.2.6) ist nicht mehr in mehreren Stunden zu leisten, wie beim Einsatz linearer Kernel, sondern erfordert mehrere Wochen Rechenzeit.

6.1.4. Die Evaluation der Wortoberflächenmerkmale und der Wortlänge

Um den Einfluss der Wortoberflächenmerkmale und der Merkmale zur Kodierung der Wortlänge (siehe 5.2.2) einzuschätzen, werden diese in Experimenten aus der Merkmalsmenge ausgeschlossen.

	LOC F-Measure	PER F-Measure	ORG F-Measure	MISC F-Measure	Gesamt Precision	Gesamt Recall	Gesamt F-Measure
Alle Merkmale (gewähltes Verfahren)	50.60	42.67	45.38	17.98	70.22%	29.28%	41.33
keine Wortoberflächen-merkmale	46.59	37.08	13.03	18.59	64.95%	20.40%	31.05
keine Merkmale für Wortlänge	50.26	41.92	45.14	16.87	70.15%	28.74%	40.77

Abbildung 6.5: Die Ergebnisse beim Verzicht auf Wortoberflächenmerkmale und auf die Merkmale zur Kodierung der Wortlänge

Wie aus den Ergebnissen in Abbildung 6.5 entnommen werden kann, tragen die Wortoberflächenmerkmale maßgeblich zur Identifizierung bei. Ein Verzicht darauf senkt das F-Measure um zehn Punkte. Dramatisch fallen vor allem die Werte bei der Kategorie ORGANISATION, was sicherlich zu großen Teilen auf das Wortoberflächenmerkmal „besteht nur aus Großbuchstaben“ zurückgeführt werden kann, welche wie bereits in 6.1.1 gezeigt, die Erkennung von Akronymen von Firmennamen unterstützt.
Das Merkmal zur Kodierung der Wortlänge hingegen zeigt nur einen verhältnismäßig geringen Einfluss, der jedoch bei allen Kategorien messbar ist.

6.1.5. Die Evaluation der erweiterten N-Gram Modellierung

In Abschnitt 5.2.1 wird neben der traditionellen N-Gram Modellierung (eingesetzt in [RÖSSLER 2004a], [RÖSSLER 2004b]) zur Erfassung eines lokalen Kontexts die erweiterte N-Gram Modellierung ([RÖSSLER & MORIK 2005]) vorgeschlagen. Dabei werden nicht mehr nur isolierte Wörter, sondern auch Wortsequenzen klassifiziert, die potentiell eine NE sind. Die erweiterte N-Gram Modellierung ist durch das Bestreben motiviert, die einzelwortbasierte „klassische" N-Gram Modellierung zu überdenken und eine Repräsentation zu entwerfen, die den Phrasen- bzw. Sequenz-Charakter von NEs sehr viel angemessener abbildet. Dadurch, dass dem Lerner nur die gesamte Sequenz als positive Instanz, Teile davon aber als negative Instanz präsentiert werden, kann eine höhere Precision mit einem möglicherweise schwächeren Recall erwartet werden. Schließlich entspricht die gewählte Repräsentation im Gegensatz zur wortbasierten Klassifikation genau dem exakten Treffer-Maß der NER-Evaluation.

Wird eine maximale Länge von beispielsweise fünf Wörtern für eine NE-Sequenz angenommen, führt dies bei konsequenter Expansion jeder Sequenz zu einer Verfünffachung der Instanzen. Da der Zeitaufwand des Lernalgorithmus der SVM abhängig von der Anzahl der Instanzen ist, ist dieser Effekt äußerst unerwünscht. In 5.2.1 sind folgende Heuristiken vorgeschlagen um die Anzahl der Instanzen zu reduzieren.

- Eine NE-Sequenz überschreitet niemals eine Satzgrenze.
- Eine NE-Sequenz endet und beginnt nicht mit kleingeschriebenen Einheiten oder Satzzeichen.
- Anhand der Trainingsdaten können „Stopp-Sequenzen" abgeleitet werden, also Wortformen und Zweiwortsequenzen, die niemals Teil einer NE sind.

In den deutschen CoNLL-Daten wurden folgende „Stopp-Sequenzen" festgelegt: Ausgehend von einer Frequenzschwelle von 110 wurden 117 Wortformen extrahiert, die niemals als NE vorkamen. Für die Bigramme wurden mit einer Frequenzschwelle von mehr als fünf Vorkommen 1194 Bigramme extrahiert, die niemals als Teil einer NE vorkamen. Der sehr viel geringere Frequenzschwellwert im Vergleich zu den Unigrammen mag erstaunen, lässt sich jedoch damit rechtfertigen, dass ein Bigramm sehr viel spezifischer, d.h. weniger mehrdeutig als ein Unigram ist.

Dieses Auslassen von Einheiten kann dazu führen, dass Wörter und Sequenzen von der Klassifikation ausgeschlossen werden, die auch als NE vorkommen können. Weil eine manuelle bzw. intellektuelle Überprüfung schwer vorstellbar ist, wurde dies anhand des Testkorpus erfolgreich überprüft.

	Anzahl Instanzen in Training /Testset	LOC F-Measure	PER F-Measure	ORG F-Measure	Gesamt Precision	Gesamt Recall	Gesamt F-Measure
„Klassische“ N-Gram Modellierung	101.810 / 25.909	50.60	42.67	45.38	69.82%	34.18%	45.90
Erweiterte N-Gram Modellierung	113.245 / 30.792	52.68	44.19	49.10	89.72%	33.13%	48.39

Abbildung 6.6: Traditionelle und erweiterte N-Gram Modellierung im Vergleich

Abbildung 6.6. zeigt nicht nur die Ergebnisse der beiden N-Gram Modellierungen, sondern auch die Anzahl der erzeugten Instanzen für das Trainings- bzw. Testset. Die Annotationen der Kategorie MISCELLANEOUS wurden für das Experiment vollständig ignoriert. Wie aus dem Vergleich der Anzahl der Instanzen deutlich wird, hat sich das vorgeschlagene Verfahren zur Reduktion der Instanzen bewährt. Darüber hinaus zeigt sich auch, dass die erwartete höhere Precision der erweiterten N-Gramm Modellierung bei allen Kategorien deutlich zu beobachten ist, während der Rückgang des Recalls erfreulicherweise nur sehr bescheiden ausfällt, so dass insgesamt ein Anstieg des F-Measure über alle Kategorien von fast vier Punkten erreicht wird.

Aufgrund der höheren Precision entsprechen die Ergebnisse der erweiterten N-Gram Modellierung sehr viel mehr den Anforderungen des Basisklassifizierer, welcher zur Erzeugung lexikalischer Ressourcen aus nicht-annotierten Daten eingesetzt wird. In den Experimenten in 6.3 wird untersucht, inwiefern sich diese Eigenschaften positiv auf die erzeugten lexikalischen Prozesse auswirken.

Als Nachteile des erweiterten N-Gram Klassifizierers ist jedoch eine erheblich geringere Geschwindigkeit anzuführen. Diese liegt kaum an der leicht erhöhten Anzahl der Instanzen, sondern am Aufwand der Sequenzextraktion und an der Rückprojektion der Sequenzklassifikationen auf den ursprünglichen Text. Darüber hinaus muss geprüft werden, ob die Verfahren zur Reduktion der Instanzen problemlos auf beliebige Korpora übertragbar

sind. Auch die auf dem Testkorpus basierende Überprüfung, ob bzw. welche Fehlerquote durch den Ausschluss von Wörtern und Sequenzen möglicherweise von vornherein in Kauf genommen wird, ist durch den meist geringen Umfang von Testkorpora nicht über alle Zweifel erhaben.

6.2. Adaptivität anhand biomedizinischer NER

Im Rahmen der JNLPBA-2004 wurde ein Shared Task zur biomedizinischen NER organisiert [KIM ET AL. 2004]. Dazu wurden Trainings- und Testdaten annotiert und allen Teilnehmern zur Verfügung gestellt (siehe auch Abschnitt 2.2.3). Abbildung 6.7 zeigt detaillierte Angaben zu den annotierten Korpora. Die Aufgabe war die ideale Herausforderung, die Korpus-Adaptivität des entwickelten Verfahrens zu evaluieren. Die Ergebnisse sind bereits in [RÖSSLER 2004a] veröffentlicht.

	Protein	DNA	RNA	Cell Type	Cell Line	Wörter
Training Set	30.269	9.533	951	6.718	3.830	472.006
Test Set	5.067	1.056	118	1.921	500	54173

Abbildung 6.7: Überblick über Trainings- und Testdaten zur biomedizinischen NER

Grundsätzlich sollte das bereits entwickelte System mit möglichst geringen Veränderungen eingesetzt werden. Dies war nicht nur ein Gebot der beabsichtigten Evaluation der Adaptivität des Verfahrens, sondern folgte schlichtweg aus der kurzen, zur Verfügung stehenden Zeit zwischen der Kenntnisnahme der Shared Task Ausschreibung und dem Abgabetermin. Das System war grundsätzlich sehr schnell für die neue Domäne einsetzbar, doch ergab sich nach der Inspektion der ersten Ergebnisse der Wunsch nach Optimierungen. Diese bezogen sich vor allem auf die korrekte Identifikation längerer NE-Sequenzen, da oft beobachtet werden konnte, dass große Teile längerer NEs korrekt klassifiziert wurden, aber das letzte Wort fehlte. Die in Abschnitt 5.2.2 vorgeschlagene erweiterte N-Gram Modellierung wurde erst zu einem späteren Zeitpunkt entwickelt und stand noch nicht zur Verfügung. Daraus resultierten die folgenden Anpassungen, die in der geringen Zeit sinnvoll und umsetzbar waren:

- **Die Integration eines Hidden Markov Modells:** Der TnT-Tagger ([BRANTS 2000]), der eigentlich für Wortarten entwickelt wurde, basiert auf einem Markov Modell. Er wurde auf demselben Trainingskorpus mit den NE-Labels trainiert. Die vom Modell berechneten Ausgabewahrscheinlichkeiten wurden für alle Klassen als Merkmal für die SVM benutzt.

Das Hidden Markov Modell wurde aufgrund seiner Sequenz-orientierten Klassifikation integriert. Grundsätzlich erscheint eine andere Reihenfolge viel versprechender als die hier verfolgte; nämlich die Linearisierung und Auflösung der Ausgaben der verschiedenen SVMs mittels Hidden Markov Modellierung. Allerdings wird dabei unter anderem eine effiziente Umwandlung des Funktionswerts $(\vec{w}\cdot\vec{x}+b)$ der SVM in Wahrscheinlichkeiten benötigt, was keine triviale Herausforderung darstellt (vgl. etwa [RÜPING 2004]). Doch wurde in der wenigen zur Verfügung stehenden Zeit keine erfolgreiche Lösung dafür gefunden.

- **Eine zusätzliche Post-Processing-Komponente:** Einem bisher mit „nicht-NE“ klassifizierten Wort wird dann die Klasse des vorhergehenden Wortes zugewiesen
  - wenn der betreffende Funktionswert $(\vec{w}\cdot\vec{x}+b)$ beim aktuellen Wort zu den drei höchsten Funktionswerten gehört
  - wenn der Wert eines zusätzlich eingeführten NE vs. „nicht-NE“-Klassifizierers das aktuelle Wort als NE klassifiziert.
- **Ein zusätzliches Wortoberflächenmerkmal:** Dieses dient zur Entdeckung von „ATCG“-Sequenzen. Eine „ATCG“-Sequenz bezeichnet eine DNA und benutzt dazu ihre Zusammensetzung aus den vier DNA-Bausteinen. Allerdings blieb das zusätzliche Merkmal ohne jeglichen Einfluss auf die Ergebnisse.

Ausgehend von der grundlegenden Konfiguration des Basisklassifizierers wurden folgende Experimente durchgeführt:

(1) Grundlegende Konfiguration des Basisklassifizierers: Substring-Repräsentation, kodierte Wortlänge und Wortoberflächenmerkmale

(2) Wie (1), aber ganze Wortformen anstelle der Substring-Repräsentation

(3) Wie (1), aber mit dem zusätzlichen Postprocessing

(4) Wie (1), aber mit den zusätzlichen Merkmalen der Ausgabewahrscheinlichkeiten des Hidden Markov Modells

(5) Wie (1), aber mit dem zusätzlichen Postprocessing und den zusätzlichen Merkmalen der Ausgabewahrscheinlichkeiten des Hidden Markov Modells

(6) Nur das Hidden Markov Modell

	Protein F-Measure	DNA F-Measure	RNA F-Measure	Cell Type F-Measure	Cell Line F-Measure	Gesamt Recall	Gesamt Precision	Gesamt F-Measure
(1) Basisklassifizierer	62.6	50.0	48.1	56.9	36.8	61.0%	56.2%	58.5
(2) Basisklassifizierer ohne Substring-Zerlegung	60.1	48.9	48.1	51.5	34.7	57.9%	54.4%	56.1
(3) Basisklassifizierer + Postprocessing	65.0	54.2	55.0	64.0	49.2	65.4%	59.9%	62.6
(4) Basisklassifizierer + Markov Modell	66.4	52.7	44.0	65.3	47.4	66.3%	60.1%	63.1
(5) Basisklassifizierer + Markov Modell + Postprocessing	67.0	55.1	46.7	66.5	48.3	**67.4**%	**60.1**%	**64.0**
(6) nur Markov Model	62.6	47.5	36.4	58.4	39.7	62.6%	54.1%	58.0

Abbildung 6.8: Experimente zur Korpus-Adaptivität des Verfahrens anhand biomedizinischer Korpora

Abbildung 6.8 zeigt die auf dem biomedizinischen Korpus erzielten Ergebnisse. Wie im Vergleich zwischen (1) und (2) deutlich wird, bewährt sich die Zerlegung in positionale Substrings auch in der biomedizinischen Domäne. Die in (3) evaluierte Post-Processing Komponente zeigt trotz ihrer Einfachheit beeindruckende Resultate. Noch besser wird das System durch die Berücksichtigung der Ausgabewahrscheinlichkeiten des Markov Modells (4). Und die beste Konfiguration integriert beide Erweiterungen und zeigt damit ein weiteres Mal die hervorragenden Eigenschaften der SVM, vielfältige Evidenzen miteinander zu kombinieren. Gleichzeitig weist das Ergebnis aber auch darauf hin, dass das aktuelle Modell deutlich von einer stärkeren Berücksichtigung des Sequenzcharakters von NEs profitiert, wenn auch eine elegante Lösung hierzu noch aussteht.

Die beste Konfiguration (5) schneidet im Vergleich mit den anderen Beiträgen des Shared Task [KIM ET AL. 2004] im unteren Mittelfeld ab und erreichte den sechsten Rang von acht Teilnehmern. Der beste Beitrag ([ZHOU & SU 2004]) erreicht ein F-Measure über alle Kategorien von 72.6, das schwächste System ([LEE ET AL. 2004]) lag bei 49.1. Damit sind die Ergebnisse des Ansatzes auch im Vergleich zu den anderen Systemen zumindest respektabel, insbesondere wenn der geringe Zeitaufwand und die Unerfahrenheit des Systementwicklers mit der Domäne in Betracht gezogen werden. Darüber hinaus kann ganz klar gezeigt werden, dass der Ansatz in hohem Maße Korpus-adaptiv ist und sich seine Eigenschaften nicht nur in anderen Domänen, sondern auch in anderen Sprachen bewähren.

Auch auf der biomedizinischen Domäne wurde mit dem im nächsten Abschnitt evaluierten Verfahren zur automatischen Erzeugung lexikalischer Ressourcen experimentiert. Doch weder die Dokumenten-orientierte Klassifikation der Postprocessing-Komponente (Abschnitt 5.4) noch der Erwerb interner Evidenz mit dem System I (Abschnitt 5.3.1) aus umfangreichen nicht-annotierten Daten konnten die bereits erreichten Ergebnisse verbessern. Als Ursache für den gescheiterten Einsatz von System I ist die schwache Precision der Erkennungsrate anzunehmen. Der mangelnde Erfolg der Dokumenten-orientierten Klassifikation ist vermutlich darauf zurückzuführen, dass die zu bearbeitenden Texte Abstracts von Aufsätzen sind. Es ist anzunehmen, dass Wiederholungen einer NE innerhalb von solchen Zusammenfassungen seltener sind als in den „Volltexten“ des FR-Corpus. Eine nähere Untersuchung der Befunde und weitere Experimente wären von großem Interesse.

6.3. NER mit den erweiterten Merkmalen

Der zentrale Punkt des hier entwickelten Korpus-adaptiven Verfahrens ist die Optimierung des in 6.1 evaluierten Basisklassifizierers durch die Anreicherung der Instanzen anhand automatisch annotierter Daten des zu verarbeitenden Korpus. Der daraus resultierende erweiterte Klassifizierer muss dem Vergleich anderer Verfahren standhalten, d.h. regelbasierten Systemen oder Systemen, die mit umfangreichen Ressourcen ausgestattet sind.
In 6.3.1 und 6.3.2 werden die mit System I (5.3.1) bzw. System II (5.3.2) erzeugten erweiterten Merkmale mit weiteren NER-Systemen verglichen. Als nicht-annotiertes Datenmaterial wurde in allen Versuchen das mehr als 40 Millionen Wörter umfassende FR-Korpus ([FR-CORPUS 1994]) eingesetzt. In 6.3.3 wird der Beitrag der Postprocessing-Komponente sowohl für die Basis-, als auch für die erweiterten Klassifizierer untersucht. Um die durchgehende Datenorientierung des Ansatzes zu evaluieren, also den Verzicht auf linguistische Generalisierung, wird in 6.3.4 der Einfluss der Kodierung von Wortarten untersucht. Dies wird sowohl am Basisklassifizierer als auch am erweiterten Klassifizierer getestet. Der Vergleich mit dem erweiterten Klassifizierer ist deshalb bedeutsam, da der Verzicht auf zusätzliche Ressourcen nur vor dem Hintergrund des Einsatzes automatisch erzeugter Ressourcen sinnvoll ist. Um auch Fragen der Verarbeitungsgeschwindigkeit angemessen zu beleuchten, sind in 6.3.5 verschieden Laufzeitmessungen beschrieben und diskutiert.

6.3.1. Die Auswertung automatisch annotierter Daten – System I

System I zur Merkmalsanreicherung der Instanzen mit interner Evidenz aus automatisch annotierten Daten (Abschnitt 5.3.1) basiert auf einer Vielzahl von Versuchen mit dem Basisklassifizierer und der Anwendung desselben auf die 40 Millionen Wörter des FR-Korpus ([FR-Corpus 1994]). Ausgehend von der Idee der idealen Liste (vgl. Abschnitt 5.3) wurden mit der Kategorie PERSON eine Vielzahl von Experimenten durchgeführt, von denen hier nur die zwei wichtigsten evaluiert werden:

- Die naive Umsetzung der Idee der idealen Liste, die gewissermaßen als Ausgangspunkt betrachtet werden kann.
- Die anschließend gewählte Umsetzung, die in vielen Experimenten zu den besten Ergebnissen führte und auch in [RÖSSLER 2004b] eingesetzt wurde.

Die Abbildung 6.9 vergleicht den naiven Ansatz mit dem auf diskretisierten Funktionswerten basierenden Verfahren. In beiden Verfahren werden die aus den nicht-annotierten Daten abgeleiteten Merkmale nicht nur für das zu klassifizierende Wort, sondern auch für die Token im Kontext erzeugt.

	Durchgänge	Gesamt Precision	Gesamt Recall	Gesamt F-Measure
Kategorie PERSONEN				
Basisklassifizierer – die Ausgangslage	-	69.23%	30.84%	42.67
Naiver Ansatz	1	78.76%	53.79%	63.92
Diskretisierte Funktionswerte (gewähltes Verfahren)	2	86.72%	67.31%	75.79
Naiver Ansatz + Postprocessing	1	78.03%	69.25%	73.28
Diskretisierte Funktionswerte (gewähltes Verfahren) + Postprocessing	2	86.77%	**91.40%**	**89.03**
[VOLK & CLEMATIDE 2001]	-	92%	86%	88,9
[NEUMANN & PISKORKSI 2002]	-	**95.9%**	81.3%	88.0
Bester CoNLL-Beitrag [KLEIN ET AL. 2003]	-	73.54%	60.65%	83.57

Abbildung 6.9: Evaluation der Kategorie PERSON zur Extraktion von Evidenz aus umfangreichen, nicht-annotierten Daten (System I)

Obwohl auch der naive Ansatz die Erkennungsrate im Vergleich zum Basisklassifizierer beeindruckend zu steigern vermag, so zeigt er sich dem gewählten Verfahren gegenüber doch eindeutig unterlegen. Trotz der Steigerung der Precision des naiven erweiterten Klassifizierers sinken die Ergebnisse in einem erneuten Durchgang ab, so dass im Gegensatz zum gewählten Verfahren nur ein Durchgang durchgeführt wird. Beim gewählten Verfahren steigert sich das Ergebnis im zweiten Durchgang allerdings nur minimal. Die Überlegenheit des Verfahrens ist auch nach der Integration der Postprocessing-Komponente uneingeschränkt erhalten.
Die Ergebnisse der Kategorie PERSON sind äußerst beeindruckend, werden doch damit alle Beiträge der CoNLL-2003 ([TJONG KIM SANG & DE MEULDER 2003]) deutlich übertroffen; gleichzeitig stehen die Resultate den beiden regelbasierten Systemen ([NEUMANN & PISKORKSI 2002], [VOLK & CLEMATIDE 2001]) bezüglich des F-Measure in nichts nach. Deswegen wurde das Verfahren genauso auf die Kategorien ORGANISATION und LOCATION angewandt.

	Durchgänge	Gesamt Precision	Gesamt Recall	Gesamt F-Measure
Kategorie ORGANISATION				
Basisklassifizierer	-	63.80%	35.21%	45.38
Erweiterter Klassifizierer	1	62.71%	48.87%	54.93
[VOLK & CLEMATIDE 2001]	-	76%	81%	78.4
[NEUMANN & PISKORKSI 2002]	-	96.7%	67.3%	79.4
Bester CoNLL-Beitrag [FLORIAN ET AL. 2003]	-	83.64%	61.80%	71.08
Kategorie LOCATION				
Basisklassifizierer	-	77.72%	37.51%	50.60
Erweiterter Klassifizierer	1	74.00 %	59.88%	66.20
[VOLK & CLEMATIDE 2001]	-	81%	91%	85.7
[NEUMANN & PISKORKSI 2002]	-	88.2%	75.1%	81.1
Bester CoNLL-Beitrag [FLORIAN ET AL. 2003]	-	83.19%	72.90%	77.71

Abbildung 6.10: Evaluation der Kategorien ORGANISATION und LOCATION zur Extraktion interner Evidenz aus umfangreichen, nicht-annotierten Daten (System I)

Abbildung 6.10 zeigt die Ergebnisse für die Kategorien ORGANISATION und LOCATION. Zwar zeigen die Werte der erweiterten Klassifizierer eine Verbesserung von 9 bzw. 16

Punkten an, liegen jedoch weiterhin deutlich unter dem besten CoNLL-Beitrag ([FLORIAN ET AL. 2003]) und den Ergebnissen der regelbasierten Ansätze ([NEUMANN & PISKORKSI 2002], [VOLK & CLEMATIDE 2001]). Der Einsatz nicht-annotierter Daten wurde bereits nach einem Durchgang beendet, da eine erneute Runde zu keiner Verbesserung, sondern zu einer Verschlechterung der Resultate führte. Auch die Postprocessing-Komponente verschlechtert die Ergebnisse. Offensichtlich enthalten die Annotationen des erweiterten Klassifizierers zu viele falsche Voraussagen, so dass daraus keine hilfreichen Ressourcen anhand der automatisch annotierten Ressourcen erzeugt werden können.

Nur kurz erwähnt sei hier die erfolglose Anwendung des Verfahrens auf die biomedizinische Domäne. In identischer Art und Weise wurde das in 6.2 beschriebene Verfahren, allerdings ohne Post-Processing, als Basisklassifizierer auf ein 100 Millionen Korpus von Texten derselben Domäne angewandt. Während die Texte des annotierten Korpus auf dem Ergebnis der MEDLINE ([MEDLINE]) Abfrage mit den Termen „human“, „blood cells“ und „transcription factors“ beruht, wurde das größere Korpus mit den nicht-annotierten Texten nur mit den Termen „blood cells“ und „transcription factors“ erzeugt. Allerdings führten die aus den automatischen Annotationen des Basisklassifizierers gewonnenen Ressourcen zu keinerlei Verbesserungen.

6.3.2. Die Auswertung automatisch annotierter Daten – System II

Die Ergebnisse von System I zur Extraktion von Evidenz aus nicht-annotierten Daten sind äußerst ermutigend. Ausgehend von den Beschränkungen von System I wurde deshalb in 5.3.2 die Erweiterung zu System II vorgeschlagen. Dabei wird zum einen die traditionelle durch die erweiterte N-Gram Modellierung (5.2.1) ersetzt, zum anderen beschränkt sich die Auswertung automatisch annotierter Daten nicht mehr auf interne Evidenz, sondern es wird simultan interne und externe Evidenz erworben.

Wie bereits in 6.1.5 untersucht, weist die erweiterte N-Gram Modellierung im Vergleich zur klassischen N-Gram Modellierung eine höhere Precision bei leicht besserem F-Measure auf. Aufgrund dieser Eigenschaften eignet sich dieser Ansatz theoretisch besser für die Erzeugung von Ressourcen aus nicht-annotierten Daten.

	Durchgang	Anzahl der Support Vektoren		
		LOC	PER	ORG
Basisklassifizierer	**0**	9579	6796	7138
Nach jedem Durchgang aktualisierte interne und externe Evidenz	**1**	7234	4765	6224
	2	6562	4027	5871
	3	6423	4015	5699
	4	6353	4002	5681

Abbildung 6.11: Anzahl der Support Vektoren im Verfahren zur Extraktion interner und externer Evidenz aus umfangreichen, nicht-annotierten Daten (System II)

Wie den Ergebnissen in Abbildung 6.12 entnommen werden kann, entsprechen die experimentellen Ergebnisse nicht dieser Erwartung. Zwar zeigt sich in der ersten Runde eine deutliche Verbesserung im Vergleich zum Basisklassifizierer, doch ist bei den weiteren Durchgängen nur noch eine geringe Steigerung zu beobachten. Diese Verbesserung betrifft interessanterweise nicht nur Recall sondern auch Precision. Darüber hinaus zeigt eine Inspektion der SVM-Modelle eine interessante Beobachtung: Obwohl die Leistung der Modelle sich nur unerheblich verbessert, werden die Modelle bei jedem Durchgang kleiner bezüglich der Anzahl der Support Vektoren (Abbildung 6. 11). Bereits in 5.3.2 wurde argumentiert, dass dabei möglicherweise spezifisches Wissen über einzelne NEs bzw. darin enthaltene Wörter und Wortsequenzen durch allgemeineres Wissen über Kontexte von NEs ersetzt werde. Auf der Annahme basierend, dass das System II vorrangig zum Erwerb von Kontextmerkmalen und das einzelwortbasierte System I besser zur Erzeugung von Merkmalen der internen Evidenz geeignet ist, wird der Erwerb von Merkmalen modifiziert, nachdem sich die Anzahl der Support Vektoren nicht mehr wesentlich weiter verringerte: Die erworbenen Kontextmerkmale werden fixiert und der Erwerb beschränkt sich auf Merkmale der internen Evidenz, welche auf automatischen Annotationen berechnet werden, die genau wie in System I auf dem Einsatz traditioneller N-Gramme basieren.

	Durchgang	N-Gramm	LOC F-Measure	PER F-Measure	ORG F-Measure	Gesamt Precision	Gesamt Recall	Gesamt F-Measure
Basisklassifizierer	**0**		55.18	64.21	50.11	83.49%	43.01%	56.77
Nach jedem Durchgang aktualisierte interne und externe Evidenz	**1**	**Erweitert**	69.92	87.08	53.67	85.93%	61.99%	72.02
	2		69.88	90.38	54.37	85.83%	64.53%	73.67
	3		71.22	90.63	54.96	85.66%	65.62%	74.31
	4'		*72.21*	*90.74*	*55.25*	*85.71%*	*66.25%*	*74.73*
	4	**Klassisch**	72.88	91.55	58.96	82.73%	70.14%	75.91
	5		74.64	91.86	62.75	82.93%	72.91%	77.59
	6		75.10	91.97	63.84	82.83%	73.79%	78.05
	7		75.94	91.09	65.36	83.69%	73.82%	78.44
[VOLK & CLEMATIDE 2001]			85.7	88.9	78.4			
[NEUMANN & PISKORKSI 2002]			81.1	88.0	79.4			
Bester CoNLL-Beiträge (PER:[KLEIN ET AL. 2003]; ORG, PER: [FLORIAN ET AL. 2003])			77.71	83.57	71.08			

Abbildung 6.12: Evaluation des Verfahrens zur Extraktion interner und externer Evidenz aus umfangreichen, nicht-annotierten Daten (System II); die Ergebnisse von 4' sind nicht in das Verfahren mit eingeflossen

Abbildung 6.12 zeigt die Evaluation dieses kombinierten Verfahrens, nach der Optimierung durch die Postprocessing-Komponente (5.4). Ausgehend vom Basisklassifizierer wurden während insgesamt drei Runden mit dem System II erweiterte Merkmale aus dem automatisch annotierten umfangreichen Korpus erzeugt. Während der Runde zwei und drei ergab sich zwar keine massive Verbesserung der gesamten Erkennungsleistung, doch sank die Anzahl an Support Vektoren (Abbildung 6.11) bei gleichzeitig leichtem Anstieg von Precision und Recall. In der vierten Runde mit den erweiterten N-Grammen, in Abbildung 6.12 als ***4'*** aufgeführt, werden diese Effekte minimal, so dass für die vierte Runde die erweiterten N-Gramme durch die klassischen N-Gramme ersetzt wurden und nur noch das einzelwortbasierte Korpus nach jedem weiteren Durchgang aktualisiert wird. Wie an den Ergebnissen zu sehen ist, ergibt sich dadurch eine weitere Verbesserung in den nächsten vier Runden und zwar um mehr als 10 Punkte F-Measure bei ORGANISATION und um fast 5 Punkte F-Measure bei LOCATION. Die Kategorie PERSON ist bereits in Runde zwei höher

als beim isolierten Einsatz von System I (Abschnitt 6.3.1). In den weiteren Runden verbessern sich die Ergebnisse minimal, sinken jedoch in der letzten Runde leicht ab.
Die Ergebnisse der modifizierten Version von System II sind sehr viel besser als mit System I (Abbildung 6.9 und 6.10). Selbst die Leistung der Kategorie PERSON wird weiter optimiert und liegt damit sogar über den Ergebnissen der beiden regelbasierten Systeme. Allerdings sind die Erkennungsraten bei den Kategorien ORGANISATION und LOCATION weiterhin unter den besten CoNLL-Beiträgen und deutlich schwächer als regelbasierte Systeme für deutsche Texte.
Die in diesem Abschnitt evaluierte Konfiguration hat sich als die erfolgreichste herausgestellt. Obwohl das Verfahren nicht in allen Bereichen eine befriedigende Leistung erzielt, ist es doch in mehrfacher Hinsicht als erfolgreich und viel versprechend zu beurteilen. Eine abschließende Diskussion des hier verfolgten Ansatzes findet in Kapitel 6.4 statt.

6.3.3. Die Evaluation des Postprocessing

Die Dokumenten-orientierte Klassifikation ist gemeinsam mit der Ausnutzung des Koordinationsphänomens als eine integrierte Postprocessing-Komponente implementiert (Abschnitt 5.4). Abbildung 6.13 zeigt die Ergebnisse von insgesamt drei Konfigurationen, jeweils mit und ohne Postprocessing-Komponente. Als Konfigurationen wurde ein Basisklassifizierer mit traditioneller N-Gram Modellierung, ein Basisklassifizierer mit erweiterter N-Gram Modellierung sowie ein erweiterter Klassifizierer (System II) eingesetzt. Der erweiterte Klassifizierer entspricht dem in 6.3.2 evaluierten Verfahren zum Einsatz nicht-annotierter Daten und stellt in Kombination mit der Postprocessing-Komponente die besten in dieser Arbeit erreichten Resultate dar.
Der Einsatz der Postprocessing-Komponente führt bei allen Konfigurationen zu erheblichen Verbesserungen des F-Measure. Allerdings wird der Einfluss schwächer, je besser die Ausgangsleistung des Systems ist. Betrachtet man die einzelnen Kategorien, so zeigen sich deutliche Unterschiede. Am stärksten profitiert die Kategorie PERSON, sehr viel schwächer nur die Kategorien ORGANISATION und LOCATION.

	Postprocessing	LOC F-Measure	PER F-Measure	ORG F-Measure	MISC F-Measure	Gesamt Precision	Gesamt Recall	Gesamt F-Measure
Basisklassifizierer mit klassischen N-Grammen		50.60	42.67	45.38	17.98	70.22%	29.28%	41.33
	X	52.92	71.77	49.00	19.19	72.93%	41.75%	53.11
Basisklassifizierer mit erweiterten N-Grammen		52.68	44.19	49.10	**Nicht evaluiert**	89.72%	33.13%	48.39
	X	55.18	64.21	50.11		83.49%	43.01%	56.77
Erweiterter Klassifizierer (System II, 6.3.2)		74.62	78.22	62.68		84.17%	63.14%	72.16
	X	75.94	91.09	65.36		83.69%	73.82%	78.44

Abbildung 6.13: Die Evaluation des Postprocessing

6.3.4. Die Evaluation der Kodierung von Wortarten

Wie in Abschnitt 5. 1 beschrieben, verzichtet unser Ansatz vollständig auf den Einsatz linguistischer Werkzeuge. Am häufigsten für die NER eingesetzt sind sicherlich die POS-Tagger, Systeme zur automatischen Zuweisung der Wortart. Um den positiven Einfluss der Kodierung von Wortarten einzuschätzen, wurde eine ganze Reihe von Experimenten jeweils mit und ohne kodierte Wortarten durchgeführt. Dies ist erforderlich, da nur über mehrere unterschiedliche Experimente ein differenziertes Bild des Einflusses von Wortarten gewonnen werden kann. Werden die Wortarten beispielsweise nur beim Basisklassifizierer vergleichend eingesetzt, so ist eine verhältnismäßig große Wirkung zu erwarten, da der Basisklassifizierer auf sehr wenigen Merkmalen operiert. Die Motivation des hier verfolgten Ansatzes ist es, die fehlenden Ressourcen und die nicht eingesetzten linguistischen Werkzeuge durch Verfahren zu ersetzen, die Evidenz zur NER aus nicht-annotierten Daten ableiten. Deshalb muss eine angemessene Evaluation zum Einfluss der Kodierung von Wortarten mit dem durch solche Ressourcen erweiterten Klassifizierer durchgeführt werden. Um ein möglichst genaues Bild vom Einfluss kodierter Wortarten zu erhalten, werden darüber hinaus alle Experimente auch mit und ohne Postprocessing (5.4) durchgeführt.

Die Experimente werden mit den POS-Tags durchgeführt, welche zusammen mit den CoNLL-Daten ausgeliefert werden. Die Tags sind vom System TreeTagger ([SCHMID 1995]) ohne nachträgliche manuelle Korrekturen eingefügt worden.

	Wortart	Klassifizierer	LOC F-Measure	PER F-Measure	ORG F-Measure	MISC F-Measure	Gesamt Precision	Gesamt Recall	Gesamt F-Measure
Klassische N-Gramme		Basis	50.60	42.67	45.38	17.98	70.22%	29.28%	41.33
	X		59.32	57.77	45.81	15.93	72.55%	36.42%	**48.49**
Klassische N-Gramme + Post-processing			52.92	71.77	49.00	19.19	72.93%	41.75%	53.11
	X		61.16	79.33	49.04	17.64	73.52%	48.42%	**58.38**
Erweiterte N-Gramme			52.68	44.19	49.10	Nicht evaluiert	89.72%	33.13%	48.39
	X		59.32	57.77	45.81		72.42%	43.51%	**54.36**
Erweiterte N-Gramme + Post-processing			55.18	64.21	50.11		83.49%	43.01%	56.77
	X		61.16	79.33	49.04		73.46%	58.34%	**65.03**
System II - interne und externe Evidenz (6.3.2)		Erweitert	74.62	78.22	62.68		84.17%	63.14%	72.16
	X		74.92	78.13	62.70		83.65%	63.56%	**72.24**
System II - interne und externe Evidenz (6.3.2) + Postprocessing			75.94	91.09	65.36		83.69%	73.82%	78.44
	X		76.33	90.76	65.53		83.41%	74.18%	**78.53**

Abbildung 6.14: Die Ergebnisse zum Einfluss der Kodierung von Wortarten; in Fett gesetzt das jeweils bessere Ergebnis der paarweisen Vergleiche

Abbildung 6.14 zeigt die Ergebnisse der Experimente, wobei jede Konfiguration einmal mit und einmal ohne kodierte Wortarten evaluiert und das höhere F-Measure jeweils fett gesetzt wurde. Auf den ersten Blick wird deutlich, dass die Konfigurationen mit Zugriff auf die Wortarten immer besser abschneiden. Insbesondere bei allen Basisklassifizierern, sowohl mit als auch ohne Postprocessing, ist ein deutlich positiver Einfluss der kodierten Wortart feststellbar. Bei den erweiterten Klassifizierern hingegen, die Zugriff auf die automatisch extrahierte interne und externe Evidenz haben, ist der positive Einfluss der Wortarten kaum mehr feststellbar und geht insbesondere auf Kosten der Precision-Werte.

Die Ergebnisse machen deutlich, dass die Kodierung von Wortarten einen positiven Einfluss auf die Erkennungsrate hat. Allerdings verschwindet dieser positive Einfluss nahezu

vollständig, wenn die Ressourcen zum Einsatz kommen, welche durch das Verfahren zur Auswertung automatisch annotierter Daten erzeugt werden.

6.3.5. Laufzeiten im Vergleich

Um eine Einschätzung des Zeitaufwandes zu ermöglichen, sind in Abbildung 6.15 und 6.16 verschiedene Laufzeitmessungen aufgeführt. Das System ist zurzeit keinesfalls für Geschwindigkeit, sondern für effizientes Debugging und Modifizieren der einzelnen Schritte optimiert. Geschwindigkeitssteigerungen sind daher durch eine Überarbeitung des Codes leicht zu erreichen. Alle Versuche wurden auf einem 2GHz Rechner mit Fedora ausgeführt, der darüber hinaus über 2GB Arbeitsspeicher verfügt. Diese große Menge an Arbeitsspeicher ist nicht für die Klassifikation erforderlich, beschleunigt jedoch die Berechnung der erweiterten Merkmale (5.3), bei denen eine Vielzahl von Klassifikationsentscheiden zu verarbeiten sind.

Aufgabe	Instanzen	Kernel	PER	ORG	LOC	Gesamt
SVM Training auf CoNLL-Daten	100'137 Instanze nur Basismerkmale	linear	111s	110s	134s	355s
		polynomial	490s	350s	809s	1698s
SVM Klassifikation der CoNLL-Test-daten	25'909 Instanzen nur Basismerkmale	linear	<3s	<3s	<3s	<9s
		polynomial	328s	387s	423s	1138s
SVM Klassifikation der CoNLL-Test-daten	25'909 Instanzen; Basis- und erweiterte Merkmale	linear	8s	9s	9s	26s
SVM Klassifikation des gesamten FR-Korpus ([FR-CORPUS 1994])	19'651'516 Instanzen; nur Basismerkmale	linear	-	-	-	<2h
		polynomial	-	-	-	~240h
	19'651'516 Instanzen; Basis- und erweiterte Merkmale	linear	-	-	-	<6h

Abbildung 6.15: Zeitaufwand für SVM-Training und Klassifikation

Abbildung 6.15 zeigt Messungen des SVM-Trainings und der Klassifikation bereits erstellter Instanzen. Der Vergleich zwischen polynomialem und linearem Kernel macht deutlich, dass nicht der Zeitaufwand des Trainings, sondern der Zeitaufwand der Klassifikation den Einsatz polynomialer Kernel erheblich erschwert, wenn nicht sogar unmöglich macht. In den Messungen zeigt sich außerdem, dass die Geschwindigkeit des SVM-Einsatzes kaum von der Anzahl der verwendeten Merkmale abhängt. Die festgestellten Unterschiede sind außerdem vorrangig auf Input-Output Vorgänge zurückzuführen.

Abbildung 6.16 zeigt den Zeitaufwand für die komplette NE-Annotation des Testkorpus der deutschen CoNLL-Daten, wobei drei Konfigurationen aufgeführt sind.

(1) Die traditionelle N-Gram Modellierung, bei der nur die Basis-Merkmale verwendet werden. Dies entspricht einem einfachen Basisklassifizierer (5.3).

(2) Die traditionelle N-Gram Modellierung, bei der die Basis- und alle erweiterten Merkmale verwendet werden. Dies entspricht dem Klassifizierer nach der abgeschlossenen Erzeugung der erweiterten Merkmale mit System II (5.3.2). Dieser Klassifizierer erreicht die besten Erkennungsraten. Der Unterschied im Laufzeitverhalten zu (1) ist einzig auf die Verwaltung und den Einsatz der erweiterten Merkmale zurückzuführen und könnte durch effizientere Programmierung sicherlich verringert werden.

(3) Die erweiterte N-Gram Modellierung, bei der die Basis- und alle erweiterten Merkmale verwendet werden. Dies entspricht einem Klassifizierer, wie er während der Erzeugung des zweiten Klassifizierers eingesetzt wird, solange mit erweiterten N-Grammen gearbeitet wird. Der deutliche Unterschied zu (2) entsteht durch die verhältnismäßig aufwändige Auswahl und Reduktion der zu verarbeitenden Mehrwortsequenzen und nicht durch eine deutlich höhere Anzahl Instanzen.

Konfiguration	**(1)**	**(2)**	**(3)**
N-Gram Modellierung	**traditionelle**		**erweitert**
Merkmale	**Basis-**	**Basis- und alle erweiterten**	
Erzeugung der 25909 Instanzen (1-3,optional 5) aus 54713 Wörter	24s	83s	158s
SVM-Klassifizierung (6)	9s	26s	32s
Tagging (7)	<2s	<2s	<2s
Postprocessing (8-9)	<2s	<2s	<2s
Gesamt	**37s**	**113s**	**194s**
(Die geklammerten Nummern beziehen sich auf die Verarbeitungsschritte in 5.4)			
Geschwindigkeit in Wörtern pro Sekunde	1479w/s	484w/s	346w/s
Anwendung auf das verwendete FR-Korpus (41866256)	<8h	~24 h	~41 h

Abbildung 6.16: Zeitaufwand für die komplette NE-Annotation einer Textsammlung

Wie an (1) zu sehen ist, ist eine Verarbeitungsgeschwindigkeit von mehr als 1000 Wörtern grundsätzlich möglich, so dass eine effizientere Programmierung auch (2) in diesen Bereich bringen könnte. Eine Verarbeitungsgeschwindigkeit von mehr als 1000 Wörtern pro Sekunde genügt für die meisten Anwendungen, da die Annotationen offline durchführbar sind und selbst ein Korpus von einer Milliarde Wörter in weniger als zwei Wochen annotierbar ist.

6.4. Diskussion der Ergebnisse

Bei der Evaluation des Basisklassifizierers konnte die gewählte Konfiguration in fast allen Bereichen positiv beurteilt werden. So wurde in 6.1.1 gezeigt, dass die gewählte Repräsentation mit positionalen Substrings gegenüber derjenigen mit der kompletten Wortform deutlich bessere Erkennungsleistungen aufweist und darüber hinaus weniger Merkmale erfordert. Unter den verschiedenen Substring-Zerlegungen sind nur die nicht-positionalen Substrings dem gewählten Verfahren ebenbürtig. Doch benötigen diese erheblich mehr Rechenaufwand. Die Evaluation des gewählten N-Gram Ausschnittes in 6.1.2 machte deutlich, dass dieser geeignet ist, externe Evidenz zu erfassen. Allerdings sind die Unterschiede der Erkennungsleistung zwischen den verschiedenen Kontextausschnitten kaum erheblich. Beim Vergleich unterschiedlicher Kernel-Funktionen in 6.1.3 hatte sich erfreulicherweise gezeigt, dass der Einsatz des sehr viel effizienteren linearen Kernels keinerlei Nachteile gegenüber dem polynomialen Kernel aufweist, sondern sogar zu besseren Ergebnissen führt. Die Experimente zum Einfluss der Wortoberflächenmerkmale in 6.1.4 zeigten, dass diese äußerst effizient zu implementierenden Merkmale von großer Relevanz für die Erkennung sind. Entgegen der Erfahrungen in den vorbereitenden Experimenten ergibt die Kodierung der Wortlänge nur einen geringen Effekt. Äußerst interessante Eigenschaften zeigte die in 5.2.1 vorgeschlagene erweiterte N-Gram Modellierung. Diese führt nicht nur zu einer deutlich höheren Precision, sondern auch zu einem insgesamt höheren F-Measure.

Die Korpus-Adaptivität des Verfahrens wurde anlässlich des Shared Task im Rahmen der JNLPBA-2004 ([KIM ET AL. 2004]) auf englischen Texten der biomedizinischen Domäne evaluiert ([RÖSSLER 2004A]). Die Ergebnisse des für diese Aufgabe leicht modifizierten Systems sind in 6.2 beschrieben worden. Obwohl das Verfahren in äußerst kurzer Zeit und ohne zusätzliche Ressourcen eingesetzt wurde, konnten im Vergleich zu den anderen Systemen respektable Ergebnisse erzielt werden. Allerdings wurde das System durch die Integration der Wahrscheinlichkeiten eines Markov Modells und eine Post-Processing Komponente zur besseren Identifikation der korrekten NE-Grenze modifiziert. Diese erfolgreichen Erweiterungen weisen deutlich auf einen zu optimierenden Aspekt des Systems hin, nämlich die korrekte Behandlung längerer NE-Sequenzen. Äußerst interessant wäre deshalb auch die Anwendung der erweiterten N-Gram Modellierung auf denselben Datensatz. Aber auch der Ersatz der regelbasierten Post-Processing Komponente durch einen lernbasierten Ansatz zur Linearisierung und Auflösung der Ausgaben der verschiedenen

SVMs erscheint vielversprechend. Trotz der erforderlichen Modifikationen zeigen die Experimente deutlich, dass der Verzicht auf Ressourcen und eine rein datenorientierte Generalisierung linguistischer Einheiten ein attraktiver Weg für Korpus-adaptive Systeme ist. Dass die Dokumenten-orientierte Klassifikation keinen Effekt zeigte, kann möglicherweise mit der Textsorte erklärt werden. So sind die zu annotierenden Abstracts kürzer als die Texte des FR-Corpus, und es kann darüber hinaus angenommen werden, dass in Zusammenfassungen die wichtigsten Bereiche des Volltextes einmal erwähnt werden und deshalb die mehrfache Erwähnung derselben NE seltener vorkommt.

Der Einsatz nicht-annotierter Daten ist einer der zentralen Beiträge dieser Arbeit und wurde in zwei Varianten, System I und System II, umgesetzt und evaluiert. Anhand der Kategorie PERSON kann klar gezeigt werden, dass ein darauf basierendes System grundsätzlich in der Lage ist, existierende lern- und regelbasierte Systeme zu übertreffen. Allerdings zeigte System I (Abschnitt 6.3.2) nur einen geringen Effekt bei den Kategorien ORGANISATION und LOCATION und keinen positiven Einfluss auf der biomedizinischen Domäne. Mit System II (Abschnitt 6.3.2) konnte bei den Kategorien ORGANISATION und LOCATION jedoch eine deutliche Verbesserung der Ergebnisse erreicht werden. Während System I auf den Erwerb einzelwortbasierter interner Evidenz beschränkt ist, setzt System II die erweiterte N-Gram Modellierung, den Erwerb von Kontexten und die stärkere Berücksichtigung des Sequenzcharakters von NEs ein. Aufgrund experimenteller Befunde wurde während der Erzeugung der Ressourcen von der erweiterten N-Gram Modellierung zur klassischen N-Gram Modellierung gewechselt, was zu einer weiteren Verbesserung der Ergebnisse führte. Ein tieferes Verständnis dieses nicht unmittelbar intuitiven Phänomens könnte die Grundlage für eine weitere Optimierung des Verfahrens bilden. Wichtige Hinweise hierzu sind sicherlich, wie in [RÖSSLER & MORIK 2005] beschrieben, aus dem Bereich des „Lernens mit unterschiedlichen Sichten auf die Daten“ („learning with multiple views“) zu erwarten. Der Einsatz unterschiedlicher Sichten wird auch beim gegenseitigen Erwerb von interner und externer Evidenz angewandt (vgl. dazu Abschnitt 4.3), ist jedoch bisher noch nie auf die Einzelwort- und Sequenzbetrachtung von NEs angewandt worden. Allerdings müssen auch Unterschiede der verschiedenen NE-Kategorien in Betracht gezogen werden. So profitiert das Verfahren sicherlich von der einfachen Syntax der Kategorie PERSON, die sich meist auf Vorname und Nachname beschränkt. Diese ermöglicht den Erwerb von Vornamen durch bereits bekannte Nachnamen und umgekehrt. NEs der Kategorie ORGANISATION können hingegen eine viel komplexere Syntax aufweisen. Auch NEs der Kategorie LOCATION

können komplex sein, vorausgesetzt dass wie im CoNLL-Korpus auch Adressangaben dazu gezählt werden. Für beide Kategorien, aber auch für die langen NEs der biomedizinischen Domäne ist ein einzelwortbasiertes, auf interne Evidenz beschränktes Verfahren ungeeignet, so dass dem Erwerb von Kontexten und der Verarbeitung ganzer Sequenzen mehr Gewicht zukommt. Allerdings hängt der Erfolg stark von der korrekten Identifikation der gesamten Sequenz ab, da falsche NE-Grenzen zu falschen Kontexten und irreführenden Sequenzen führen.

Ausgehend von den schwachen Ergebnissen der Kategorie ORGANISATION und LOCATION muss jedoch gefragt werden, ob der Leistungsfähigkeit eines Ansatzes, der vorrangig auf dem Einsatz nicht-annotierter Daten beruht, nicht auch eine gewisse Grenze gesetzt ist. So existieren Ortsnamen mit geringer oder gar keiner internen Evidenz wie etwa „Zimmern“ oder „Wohlen“. Kommen diese im nicht-annotierten Korpus nur selten und nicht in stark prädiktiven Kontexten vor, so wird das System nicht in der Lage sein, diese richtig zu erkennen. Ähnliches gilt auch für Organisationsnamen wie die meist nur „Treuhand“ genannte „Treuhandanstalt“. Möglicherweise ist für einige Kategorien aufgrund solcher NEs eine hohe Erkennungsleistung nur über den zusätzlichen Einsatz von Listen möglich.

Erfolgreich zeigt sich die Evaluation der Postprocessing-Komponente auf dem FR-Corpus. Die Experimente in 6.3.1 machten deutlich, dass die dabei durchgeführte Dokumenten-orientierte Klassifikation, aber auch die Ausnutzung von Koordinationsphänomenen einen großen Beitrag zur erfolgreichen NER leistet. Allerdings ist die Komponente zur Aufdeckung und Annotation koordinierter NEs Einzelsprachen- und vermutlich auch Korpus-spezifisch und ist auf eine manuelle Anpassung an ein neues Korpus angewiesen. Es ist jedoch denkbar, während der Anpassung eines NER-Systems gezielt nach Koordinationsstrukturen zu fahnden und diese anschließend zu implementieren. In Abhängigkeit von Eigenschaften der zu verarbeitenden Texte kann der Umgang mit koordinierten Einheiten einen elementaren Beitrag zur NER darstellen, da koordinierte Einheiten keinen herkömmlichen, durch N-Gramme modellierbaren Kontext haben (vgl. dazu 3.2.3).

Eine stärkere Integration linguistischer Modellierung in der Form existierender Tools scheint wenig Erfolg versprechend. In 6.3.4 konnte gezeigt werden, dass die Integration von Wortarten dem System keine Unterstützung zur Verfügung stellt, die über die Evidenz aus den automatisch erzeugten Ressourcen hinausgeht. Zwar sind keine Experimente mit Chunkern durchgeführt worden, die möglicherweise eine korrekte Sequenzidentifikation unterstützen, doch schadete der Einsatz solcher Tools der effizienten Korpus-Adaptivität des

Verfahrens, die in beeindruckender Art und Weise am Beispiel der englischsprachigen biomedizinischen NER-Aufgabe gezeigt werden konnte.

Das Laufzeitverhalten (6.3.5) kann grundsätzlich als zufriedenstellend bzw. aussichtsreich beurteilt werden. Angesichts der Geschwindigkeit des Basisklassifizierers und den vielen Möglichkeiten zur Effizienzsteigerung scheint eine höhere Verarbeitungsgeschwindigkeit erreichbar zu sein. Wird die Grenze von tausend Wörtern pro Sekunde erreicht, was realistisch erscheint, so kann das System in punkto Geschwindigkeit als effizient einsetzbar bezeichnet werden.

7. Zusammenfassung und Ausblick

Das im Rahmen dieser Arbeit vorgestellte Dissertationsprojekt beschreibt ein Korpus-adaptives Verfahren zur Named Entity Recognition (NER), also ein NER-Verfahren, welches effizient für neue Korpora und Anwendungen einsetzbar ist.

Um den Aspekt der Korpus-Adaptivität in den Forschungskontext einzuführen, wurden eingangs Grundlagen erarbeitet. Dazu gehört die Diskussion des Begriffs der Named Entities (NEs), dessen Unschärfe unter dem Rückgriff auf linguistische Kategorien und die Analyse NE-annotierter Korpora aufgezeigt wurde. Ausgehend von den Schwierigkeiten und möglichen Lösungen der NER-Aufgabe wurde der Forschungsstand zur NER dargestellt, der die Entwicklung Korpus-adaptiver Systeme attraktiv und aussichtsreich macht.

Nach der intensiven Diskussion der Korpus-Adaptivität existierender Verfahren bzw. der Aspekte, die der Korpus-Adaptivität förderlich oder hinderlich sind, wurde als Grundriss ein Korpus-adaptives System mit folgenden Eigenschaften vorgeschlagen:

- Eine lernbasierte Modellierung,
- der weitestgehende Verzicht auf manuell erstellte lexikalische Ressourcen,
- der Verzicht auf den Einsatz externer linguistischer Werkzeuge, was zu einer durchgehend datenorientierten Generalisierung sprachlicher Einheiten führt.

Die Implementation des Vorschlages setzt lineare Support Vektor Maschinen als Lernalgorithmus ein und wendet ein Verfahren zur Extraktion von Wissen aus nicht-annotierten Daten an. Sowohl bei der datenorientierten Repräsentation als auch beim Lernen aus nicht-annotierten Daten werden Verfahren eingesetzt, die im Rahmen dieses Dissertationsprojekts entwickelt wurden.

Anhand der automatischen Annotierung von Personennamen in deutschen Texten konnte die Leistungsfähigkeit des Ansatzes gezeigt werden. Nicht nur übertraf er alle lernbasierten Systeme auf demselben Datensatz, die Resultate waren sogar vergleichbar oder leicht besser als diejenigen der beiden regelbasierten Systeme für deutsche Texte, die auf dem Einsatz umfangreicher, manuell erstellter Ressourcen beruhen. Darüber hinaus konnte durch den erfolgreichen Einsatz des Systems zur Annotation biomedizinischer NEs in englischen Texten demonstriert werden, dass das Verfahren weder an eine bestimmte Sprache noch an eine bestimmte Domäne gebunden ist. Die Ergebnisse für die Annotation der Kategorien Ortsangaben und Organisationen in deutschen Texten sind zwar ermutigend, zeigen jedoch

genau wie die Experimente in der biomedizinischen Domäne, dass das Verfahren zum Lernen aus nicht-annotierten Daten weiterer Optimierungen bedarf.

Nicht alle möglichen Aspekte einer Aufgabenstellung können im Rahmen eines zeitlich begrenzten Dissertationsprojekts verfolgt werden. Deswegen wird an dieser Stelle ein Blick auf die weitere Entwicklung des in dieser Arbeit vorgestellten Ansatzes zur Korpus-adaptiven NER geworfen.

Eine zentrale Aufgabe liegt in einem verbesserten Umgang mit dem Mehrwort-Charakter vieler NEs. Die Experimente mit der erweiterten N-Gram Modellierung sind ein erster Schritt hierzu. Weitere Experimente, insbesondere durch die Anwendung auf andere Korpora, aber auch zum Zusammenspiel zwischen der erweiterten und der klassischen N-Gram Modellierung erscheinen interessant. Alternativ können Verfahren entwickelt werden, die die Sequenzialität direkt in das Lernverfahren integrieren. Dazu gehört sicherlich die Linearisierung und Auflösung der Ausgaben der verschiedenen SVMs mittels Markov Modellierung, wozu effiziente Verfahren zur Umwandlung des Funktionswerts der SVM in Wahrscheinlichkeiten erforderlich sind. Eine weitere Richtung besteht im Ersatz der SVM durch Conditional Random Fields ([LAFFERTY ET AL. 2001]), welche direkt nach der wahrscheinlichsten Sequenz von Labeln zu einer Sequenz von Beobachtungen suchen. Von einer verbesserten Sequenzerkennung würde nicht nur die eigentliche Annotation profitieren, sondern auch die Ableitung von Wissen aus nicht-annotierten Daten, weil das Lernen aus nicht-annotierten Daten stark unter unvollständig erkannten Sequenzen leidet.

Auch die benötigten manuellen Ressourcen verdienen eine intensivere Untersuchung. Das aktuelle Verfahren, das nur auf einem annotierten Korpus basiert, nutzt den menschlichen Einsatz möglicherweise nicht optimal aus. So wird zurzeit an attraktiven Ansätzen gearbeitet, die durch einen frühen Einsatz von Lernverfahren die manuelle Annotation unterstützen ([BECKER ET AL. 2005]). Auch die Ausbeutung semi-strukturierter Ressourcen wie etwa der Online Enzyklopädie Wikipedia ([WIKIPEDIA]) wie jüngst in [TORAL & MUÑOZ 2006], kann den Bedarf an manuellen Ressourcen senken. Außerdem verbindet sich eine solche Ausrichtung ideal mit den hier erarbeiteten Methoden zum Einsatz nicht-annotierter Daten. Darüber hinaus kann überlegt werden, existierende Listen in den Prozess zu integrieren. Diese können mit menschlicher Überwachung dazu eingesetzt werden, weitere Beispiele zu annotieren und dabei gleichzeitig die Listeneinträge zu filtern und zu evaluieren. Das hier skizzierte Verfahren könnte die manuelle Annotation eines gegebenen Korpusausschnittes

durch die computer- und wissensgestützte Erzeugung von Beispielen ersetzen. Ein unterstützender Einsatz von Listen ist darüber hinaus möglicherweise hilfreich, um für die offensichtlich schwierigeren Kategorien ORGANISATION und LOCATION hohe Erkennungsleistungen zu erreichen.

Darüber hinaus ist die konkrete Anwendung von NER-Verfahren von großem Interesse, da dabei nicht nur die Erkennungsleistung, sondern auch die Kategorien-Definitionen direkt von Anwendern kritisch beurteilt und gesteuert werden. Interessant sind vor allem Anwendungen, in denen die NER nicht Teil einer umfassenderen Textanalyse ist, sondern in denen die Ausgabe eines NER-Systems einem menschlichen Anwender direkt zur Unterstützung seines Informationsbedürfnisses zur Verfügung steht. Denkbar ist die Verschlagwortung von Dokumenten, das Passagenretrieval oder die Unterstützung beim Browsen durch Dokumentensammlungen. Dieser sehr direkte Einsatz der Ergebnisse der NER dient auch der Schärfung des NE-Begriffs, da im produktiven Einsatz die Anforderungen an ein Verfahren sehr deutlich sichtbar werden.

Anhang – ausführliche Resultate zu den Experimenten in Kapitel 6

	LOC			PER			ORG			MISC			Alle		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Abschnitt 6.1.1 Evaluation der gewählten Substring-Repräsentation (P: Precision; R: Recall; F: F-Measure)															
(1) Substrings, positional (gewähltes Verfahren)	77.72%	37.51%	50.60	69.23%	30.84%	42.67	63.80%	35.21%	45.38	75.74%	10.20%	17.98	70.22%	29.28%	41.33
(2a) Wortform, Frequenz>1	79.17%	38.61%	51.91	53.59%	17.56%	26.45	70.47%	31.35%	43.39	66.67%	1.39%	2.72	68.72%	22.86%	34.31
(2b) Wortform, Frequenz>0	79.48%	39.03%	52.36	54.72%	18.20%	27.32	70.38%	31.59%	43.60	70.00%	1.39%	2.72	69.13%	23.22%	34.76
(3) Bigramm, positional	77.78%	34.97%	48.25	60.74%	30.26%	40.40	65.62%	32.15%	43.16	85.94%	5.45%	10.24	67.91%	26.71%	38.34
(4) Trigramm, nicht positional	77.10%	38.78%	51.61	67.85%	35.55%	46.65	63.61%	35.21%	45.33	77.36%	8.12%	14.70	69.54%	30.52%	42.42
(5) Bi- und Trigram, nicht positional	78.01%	38.44%	51.50	65.23%	34.69%	45.29	64.03%	34.57%	44.90	77.78%	7.62%	13.89	68.99%	29.92%	41.74
(6) Bigramm, nicht positional	74.95%	34.21%	46.98	58.53%	30.62%	40.21	65.33%	31.43%	42.44	75.00%	2.08%	4.05	65.58%	25.74%	36.97
(7) keine Repräsentation der Wortform (Deutsch)	66.67%	1.02%	2.00	0.00%	0.00%	0.00	62.04%	27.40%	38.01	0.00%	0.00%	0.00	61.43%	7.28%	13.02
(8) keine Repräsentation der Wortform (Englisch)	77.61%	31.14%	44.44	66.17%	40.66%	50.37	58.87%	32.66%	42.01	100.00%	0.11%	0.22	67.33%	29.62%	41.14
Abschnitt 6.1.2 Evaluation des gewählten N-Gram Kontexts															
Kontext-Window drei davor, zwei dahinter (gewähltes Verfahren)	77.72%	37.51%	50.60	69.23%	30.84%	42.67	63.80%	35.21%	45.38	75.74%	10.20%	17.98	70.22%	29.28%	41.33
Kontext-Window zwei davor, zwei dahinter	75.97%	36.41%	49.23	68.77%	31.12%	42.85	64.96%	35.70%	46.07	75.00%	10.10%	17.80	69.92%	29.20%	41.19
Kontext-Window zwei davor, drei dahinter	77.27%	35.99%	49.10	68.27%	30.41%	42.07	64.51%	35.29%	45.63	73.39%	9.01%	16.05	69.80%	28.55%	40.53
Kontext-Window, kein Kontext	55.52%	43.86%	49.01	38.52%	21.91%	27.93	35.26%	17.16%	23.09	73.47%	17.82%	28.69	47.23%	25.20%	32.87

	LOC			PER			ORG			MISC			Alle		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Abschnitt 6.1.3 Die Evaluation des SVM-Kernels (P: Precision; R: Recall; F: F-Measure)															
Linearer Kernel (gewähltes Verfahren)	77.72%	37.51%	50.60	69.23%	30.84%	42.67	63.80%	35.21%	45.38	75.74%	10.20%	17.98	70.22%	29.28%	41.33
Polynomialer Kernel	87.07%	27.94%	42.31	70.58%	24.48%	36.35	65.95%	29.81%	41.07	81.67%	4.85%	9.16	73.49%	22.59%	34.56
Abschnitt 6.1.4 Die Evaluation der Wortoberflächenmerkmale und der Wortlänge															
Alle Merkmale (gewähltes Verfahren)	77.72%	37.51%	50.60	69.23%	30.84%	42.67	63.80%	35.21%	45.38	75.74%	10.20%	17.98	70.22%	29.28%	41.33
keine Wortoberflächen-merkmale	64.66%	36.41%	46.59	69.07%	25.34%	37.08	49.73%	7.49%	13.03	71.05%	10.69%	18.59	64.95%	20.40%	31.05
keine Merkmale für Wortlänge	77.94%	37.09%	50.26	68.56%	30.19%	41.92	64.19%	34.81%	45.14	75.00%	9.50%	16.87	70.15%	28.74%	40.77
Abschnitt 6.1.5 Die Evaluation der erweiterten N-Gram Modellierung															
„Klassische“ N-Gram Modellierung	77.72%	37.51%	50.60	69.23%	30.84%	42.67	63.80%	35.21%	45.38	nicht evaluiert			69.82%	34.18%	45.90
Erweiterte N-Gram Modellierung	88.93%	37.43%	52.68	93.78%	28.91%	44.19	86.91%	34.22%	49.10	nicht evaluiert			89.72%	33.13%	48.39

	LOC			PER			ORG			MISC			Alle		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Abschnitt 6.3.3 Die Evaluation des Postprocessing und 6.3.4 Die Evaluation der Kodierung von Wortarten (P: Precision; R: Recall; F: F-Measure)															
Basisklassifizierer: Klassische N-Gramme	77.72%	37.51%	50.60	69.23%	30.84%	42.67	63.80%	35.21%	45.38	75.74%	10.20%	17.98	70.22%	29.28%	41.33
Basisklassifizierer: Klassische N-Gramme mit kodierter Wortart	79.91%	47.16%	59.32	72.44%	48.04%	57.77	64.71%	35.46%	45.81	75.00%	8.91%	15.93	72.55%	36.42%	48.49
Basisklassifizierer: Klassische N-Gramme + Postprocessing	75.51%	40.73%	52.92	78.82%	65.88%	71.77	61.95%	40.53%	49.00	75.51%	10.99%	19.19	72.93%	41.75%	53.11
Basisklassifizierer: Klassische N-Gramme + Postprocessing mit kodierter Wortart	76.82%	50.80%	61.16	77.39%	81.37%	79.33	62.85%	40.21%	49.04	74.81%	10.00%	17.64	73.52%	48.42%	58.38
Basisklassifizierer: Erweiterte N-Gramme	88.93%	37.43%	52.68	93.78%	28.91%	44.19	86.91%	34.22%	49.10				89.72%	33.13%	48.39
Basisklassifizierer: Erweiterte N-Gramme mit kodierter Wortart	79.91%	47.16%	59.32	72.44%	48.04%	57.77	64.71%	35.46%	45.81				72.42%	43.51%	54.36
Basisklassifizierer: Erweiterte N-Gramme + Postprocessing	83.39%	41.24%	55.18	91.35%	49.50%	64.21	74.13%	37.84%	50.11				83.49%	43.01%	56.77
Basisklassifizierer: Erweiterte N-Gramme mit kodierter Wortart + Postprocessing	76.82%	50.80%	61.16	77.39%	81.37%	79.33	62.85%	40.21%	49.04				73.46%	58.34%	65.03
Erweiterter Klassifizierer (System II)	81.82%	68.59%	74.62	93.37%	67.31%	78.22	76.15%	53.26%	62.68	nicht evaluiert			84.17%	63.14%	72.16
Erweiterter Klassifizierer (System II) + kodierte Wortarten	82.06%	68.92%	74.92	91.89%	67.95%	78.13	75.71%	53.51%	62.70				83.65%	63.56%	72.24
Erweiterter Klassifizierer (System II) + Postprocessing	81.68%	70.96%	75.94	92.81%	89.44%	91.09	73.39%	58.90%	65.36				83.69%	73.82%	78.44
Erweiterter Klassifizierer (System II) + kodierte Wortarten + Postprocessing	81.46%	71.80%	76.33	91.82%	89.72%	90.76	73.84%	58.90%	65.53				83.41%	74.18%	78.53

Abschnitt 6.3.2 Die Auswertung automatisch annotierter Daten – System II (P: Precision; R: Recall; F: F-Measure)														
	Durchgang	N-Gramm	LOC			PER			ORG			Alle		
			P	R	F	P	R	F	P	R	F	P	R	F
Basisklassifizierer	**0**	**Erweitert**	83.39%	41.24%	55.18	91.35%	49.50%	64.21	74.13%	37.84%	50.11	83.49%	43.01%	56.77
Nach jedem Durchgang aktualisierte interne und externe Evidenz	**1**		81.35%	61.30%	69.92	94.29%	80.89%	87.08	76.88%	41.22%	53.67	85.93%	61.99%	72.02
	2		81.26%	61.30%	69.88	94.29%	86.79%	90.38	75.90%	42.35%	54.37	85.83%	64.53%	73.67
	3		80.93%	63.59%	71.22	94.32%	87.22%	90.63	75.89%	43.08%	54.96	85.66%	65.62%	74.31
	4'		81.03%	65.11%	72.21	94.40%	87.36%	90.74	76.02%	43.40%	55.25	85.71%	66.25%	74.73
	4	**Klassisch**	80.04%	66.89%	72.88	93.55%	89.63%	91.55	69.63%	51.13%	58.96	82.73%	70.14%	75.91
	5		79.58%	70.28%	74.64	93.65%	90.13%	91.86	71.55%	55.88%	62.75	82.93%	72.91%	77.59
	6		79.55%	71.13%	75.10	93.73%	90.27%	91.97	71.53%	57.65%	63.84	82.83%	73.79%	78.05
	7		81.68%	70.96%	75.94	92.81%	89.44%	91.09	73.39%	58.90%	65.36	83.69%	73.82%	78.44

Abschnitt 6.3.2 Anzahl der Support Vektoren während der Extraktion interner und externer Evidenz aus automatisch annotierten Daten – System II	Durchgang	NGram	Anzahl der Support Vektoren		
			LOC	PER	ORG
Basisklassifizierer	**0**	**Erweitert**	9579	6796	7138
Nach jedem Durchgang aktualisierte interne und externe Evidenz	**1**		7234	4765	6224
	2		6562	4027	5871
	3		6423	4015	5699
	4'		6353	4002	5681
	4	**Klassisch**	6780	3479	7539
	5		6295	3239	6992
	6		6259	3235	6903
	7		6253	3219	6883

Literaturverzeichnis

[ABERDEEN ET AL. 1995] Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P., Vilain, M. (1995). *MITRE: Description of the Alembic System as Used for MUC-6*. In: [MUC-6 1995].

[ACE ANNOTATION GUIDELINES 2004] *Annotation Guidelines for Entity Detection and Tracking (EDT)*. Version 4.2.6 200400401. Linguistic Data Consortium. Nur Online verfügbar: http://www.ldc.upenn.edu/Projects/ACE/docs/EnglishEDTV4-2-6.PDF [24.4.2006].

[AI ET AL. 2003] Ai, E., Chernuchin, D., Doukkali A., Ernst-Gerlach, A., Gendera, O., Günnewig, M., Hennig, S., Ho, T., Hüsken, P., Nouri, N., Wong, H.-M. (2003). *Wissensextraktion und -fusion am Beispiel eines Kunst-Informationssystems*. Endbericht der Projektgruppe 408. Universität Dortmund. Nur Online verfügbar: http://ls6-www.cs.uni-dortmund.de/bib/fulltext/world/PG408:03.pdf [15.5.2005].

[ALTUN ET AL. 2003] Altun, Y., Tsochantaridis, I., Hofman, T. (2003). *Hidden Markov Support Vector Machines*. In: *Proceedings of the International Conference on Machine Learning (ICML-2003)*. Washington, D.C.

[APPELT & ISRAEL 1999] Appelt, D. E., Israel, J.D. (1999). *Introduction to Information Extraction Technology - A Tutorial Prepared for IJCAI-99*. Stockholm, Sweden. Online: http://www.ai.sri.com/~appelt/ie-tutorial/ [24.4.2006].

[APPELT ET AL. 1995] Appelt, D.E., Hobbs, J.R., Bear, J., Israel, D., Kameyama, M., Kehler, A., Martin, D., Myers, K., Tyson, M. (1995). *SRI International FASTUS System MUC-6 Test Results and Analysis*. In: [MUC-6 1995].

[BANKO & BRILL 2001] Banko, M. und Brill E. (2001). *Scaling to Very Very Large Corpora for Natural Language Disambiguation*. In: *Proceedings of ACL 2001*. Toulouse, France. S. 26-33.

[BAUER 1995] Bauer, G. (1996). *Übergangsformen zwischen Eigennamen und Gattungsnamen.* In: [EICHLER ET AL. 1995], S. 1616-1621.

[BAUER 1998] Bauer, G. (1998). *Deutsche Namenskunde*. 2. überarbeitete Auflage. In: *Germanistische Lehrbuchsammlung, Band 21*. Roloff H.-G. (Hrsg). Weidler Buchverlag. Berlin.

[BECKER ET AL. 2005] Becker, M., Hachey, B., Alex, B., Grover, C., (2005). *Optimising Selective Sampling for Bootstrapping Named Entity Recognition*. In: *Proceedings of the ICML-2005 Workshop on Learning with Multiple Views*, Bonn, Germany.

[BENNETT ET AL. 1997] Bennett, S.W., Aone, C., Lovell, C. (1997). *Learning to Tag Multilingual Texts Through Observation*. In: *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*.

[BERGER ET AL. 1996] Berger, A., Della Pietra, S., Della Pietra, V.(1996). *A Maximum Entropy Approach to Natural Language Processing.* In: *Computational Linguistics*, 22(1). S. 39-71.

[BERTINO ET AL. 1999] Bertino, E., Black, B., Brasher, A., Candela, V., Catania, B., Deavin, D., Di Pace, L., Esposito, F., Leo, P., McNaught, J., Persidis, A., Rinaldi, F., Semeraro, G., Zarri, G.P. (1999). *CONCERTO - Conceptual Indexing Querying and Retrieval of Digital Documents*. In: *Proceedings of the International Conference on Multimedia Computing and Systems*. Florence. S. 1106-1109.

[BIEMANN ET AL. 2003] Biemann, C., Quasthoff, U., Böhm, K., Wolff, C. (2003). *Automatic Discovery and Aggregation of Compound Names for the Use in Knowledge Discovery*. In: *Journal of Universal Computer Science*. Volume 9(6). S. 530-541.

[BIKEL ET AL. 1997] Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R. (1997). *Nymble: a High-Performance Learning Name-finder*. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP).* Washington, D.C. S. 194-201.

[BIKEL ET AL. 1999] Bikel D.M., Schwartz, R., Weischedel, R.M. (1999). *An Algorithm that Learns What's in a Name*. In: *Machine Learning (Special Issue on NLP).* 1999.

[BLACK ET AL. 1998] Black W.J., Rinaldi, F., Mowatt, D. (1998). *FACILE: Description of the NE System used for MUC-7*. In: [MUC-7 1998].

[BOGUAREV & PUSTEJOVSKY 1996] Boguraev, B. und Pustejovsky J. (Hrsg.). *Corpus Processing for Lexical Acquisition*. In: *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*. June 1993, Columbus, Ohio. MIT Press, Cambridge, MA.

[BORTHWICK 1999] Borthwick, A. (1999). *A Maximum Entropy Approach to Named Entity Recognition*. Dissertation. New York University.

[BORTHWICK ET AL. 1998] Borthwick, A., Sterling, J., Agichtein, E., Grishman, R. (1998). *NYU: Description of the MENE Named Entity System as Used in MUC-7*. In: [MUC-7 1998].

[BRANTS 2000] Brants, T. (2000). *TnT -- A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP)*. Seattle, WA.

[BRANTS 2003] Brants, T. (2003). *Natural Language Processing in Information Retrieval*. In: *Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands*. Antwerp, Belgium.

[BRILL 1992] Brill, E. (1992). *A Simple Rule-Based Part of Speech Tagger*. In: *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing.* Trento. S. 152-155.

[BURGES 1998] Burges, C. (1998). *A Tutorial on Support Vector Machines für Pattern Recognition.* In: *Data Mining and Knowledge Discovery,* 2(2). S. 121-167.

[BUßMANN 1990] Bußmann, H. (1990). *Lexikon der Sprachwissenschaft.* 2., völlig neu bearbeitete Auflage. Stuttgart, Kröner Verlag.

[CHEN & ROSENFELD 1999] Chen, S.F. und Rosenfeld, R. (1999). *A Gaussian Prior for Smoothing Maximum Entropy Models*. Technical Report. Carnegie Mellon University.

[CHINCHOR & ROBINSON 1998] Chinchor, N. und Robinson, P. (1998). *MUC-7 named entity task definition.* Version 3.5. In: [MUC-7 1998].

[CHRISTIANINI & SHAWE-TAYLOR 2000] Christianni, N. und Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press.

[CHURCH 1988] Church, K. (1988). *A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text*. In: *Proceedings of the Second Conference on Applied Natural Language Processing*. Austin, Texas.

[COATES-STEPHENS 1991] Coates-Stephens, S. (1991). *Automatic lexical acquisition using within-text descriptions of proper nouns*. In: *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*. S. 154-169.

[COLLINS & SINGER 1999] Collins, M. und Singer, Y. (1999). *Unsupervised Models for Named Entity Classification*. In: *Proceedings of the Joint SIGDAT Conference on*

Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC 99). Maryland. S. 90-99.

[CoNLL 2002] *Proceedings of the Conference on Natural Language Learning (CoNLL-2002)*. Taipei, Taiwan. Morgan Kaufmann.

[CoNLL 2003] *Proceedings of the Conference on Natural Language Learning (CoNLL-2003)*. Edmonton Canada. Morgan Kaufmann.

[Cucchiarelli & Velardi 1999] Cucchiarelli, A. und Velardi, P. (1999). *Adaptability of linguistic resources to new domains: an experiment with proper noun dictionaries*. In: *Proceedings of the Vextal Conference*. Venice, Italy. S. 25-30.

[Cucerzan & Yarowsky 1999] Cucerzan, S. und Yarowsky, D. (1999). *Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence*. In: *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC 99)*. Maryland. S. 132-138.

[Cucerzan & Yarowsky 2002] Cucerzan, S. und Yarowsky, D. (2002). *Language independent NER using a unified model of internal and contextual evidence*. In: [CoNLL 2002]. S. 171-174.

[Cunningham 2000] Cunningham, H. (2000). *Software Architecture for Language Engineering*. Ph.D. thesis. University of Sheffield. Nur Online verfügbar: http://gate.ac.uk/sale/thesis/ [24.4.2006].

[Cunningham et al. 2003] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Ursu, C., Dimitrov, M. (2003). *Developing language processing components with GATE (a user guide)*. Nur Online verfügbar: http://gate.ac.uk/sale/tao/index.html [24.4.2006].

[Daelemans et al. 2003] Daelemans, W., Zavrel, J., van der Sloot, K. (2003). *TiMBL: Tilburg Memory-Based Learner -Version 5.0, Reference Guide*. ILK Technical Report - ILK 03-10. Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University.

[De Meulder & Daelemans 2003] De Meulder F. und Daelemans, W. (2003). *Memory-Based Named Entity Recognition using Unannotated Data*. In: [CoNLL 2003]. S. 208-211.

[DEMETRIOU & GAIZAUSKAS 2003] Demetriou G, Gaizauskas R. (2003). *Corpus resources for development and evaluation of a biological text mining system*. In: *Proceedings of the Third Meeting of the Special Interest Group on Text Mining*. Brisbane, Australia. Nur Online: http://www.pdg.cnb.uam.es/BioLink/SpecialInterestTextMining /PRESENTATIONS/rob_g.ppt [24.4.2006]

[DEMPSTER ET AL. 1977] Dempster, A.P., Laird. N.M., Rubin, D.B.(1977). *Maximum-Likelihood from incomplete data via the EM algorithm*. In: *Journal of the Royal Statistical Society*, Series B, 39(1). S. 1–38.

[DFKI-REGISTRY] The Natural Language Software Registry. Bereitgestellt und gepflegt vom DFKI in Saarbrücken. http://registry.dfki.de/. [24.4.2006].

[DUDEN 1998] *Duden Grammatik der deutschen Gegenwartssprache* (1998). 6. Auflage. Herausgegeben von der Dudenredaktion. Bearbeitet von Eisenberg, P., Gelhaus, H., Henne, H., Sitta, H., Wellmann, H. Duden Band 4. *Der Duden in 10 Bänden. Das Standardwerk zur deutschen Sprache*. Mannheim, Duden.

[EICHLER ET AL. 1995] *Namenforschung - Ein internationales Handbuch zur allgemeinen und europäischen Onomastik*. Eichler, E., Hilty, G., Löffler, H., Steger, H., Zgusta, L. (Hrsg.). De Gruyter. Berlin, New York.

[EISENBERG 1989] Eisenberg, P. (1989). *Grundriß der deutschen Grammatik*. 2. überarbeitete und erweiterte Auflage. Stuttgart, Metzler.

[FELLBAUM 1998] Fellbaum, C. (Hrsg.) (1998). *WordNet: an electronic lexical database*. Cambridge, MIT Press.

[FILLMORE 1992] Fillmore, C. J. (1992). *"Corpus linguistics" or "Computer-aided armchair linguistics"*. In: *Directions in Corpus Linguistics*. Svartvik, J. (Hrsg.). Proceedings of Nobel Symposium 82, Stockholm. Berlin, New York. Mouton de Gruyter. S. 35-60.

[FINKEL ET AL. 2004] Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Sinclair, G., Manning, C. (2004). *Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web*. In: [JNLPBA-2004]. S. 88-91.

[FLORIAN ET AL. 2003] Florian, R., Ittycheriah, A., Jing H., Zhang, T. (2003). *Named Entity Recognition through Classifier Combination*. In: [CONLL 2003]. S. 168-171.

[FR-CORPUS 1994] *Frankfurter Rundschau Corpus* (1994). Published on the ECI Multilingual Text CD. Distributed by the Linguistic Data Consortium. LDC Catalog Number LDC94T5.

[FREGE 1892] Frege, G. (1892). *Über Sinn und Bedeutung*. In: *Zeitschrift für Philosophie und philosophsiche Kritik.* NF 100, 1892, S. 25-50.

[FREUND & SCHAPIRE 1997] Freund, Y und Schapire, R.E. (1997). *A decision-theoretic generalization of on-line learning and an application to boosting*. In: *Journal of Computer and System Sciences*, 55(1). S. 119–139.

[FUKUMOTO ET AL. 1998] Fukumoto, J., Masui, F., Shimohata, M., Sasaki, M. (1998). *Description of the Oki System as Used for MUC-7*. Oki Electric Industry Co., Ltd., Osaka (Japan). 1998. In: [MUC-7 1998].

[GALE ET AL. 1992] Gale, W. A., Church, K. W., Yarowsky, D. (1992). *One sense per discourse*. In: *Proceedings of DARPA speech and Natural Language Workshop*. Harriman, NY. 1992.

[GARIGLIANO ET AL.1998] Garigliano, R., Urbanowicz, A., Nettleton, D. J. (1998). *University of Durham: Description of the LOLITA system as used in MUC-7*. In: [MUC-7 1998].

[GIMÉNEZ & MÀRQUEZ 2003] Giménez, J. und Màrquez, L. (2003). *Fast and Accurate Part-of-Speech Tagging*. *The SVM Approach Revisited*. In: *Proceedings of Recent Advances in Natural Language Processing (RANLP – 2003)*. Borovets, Bulgary. John Benjamins Publishers. S. 158 – 165.

[GRISHMAN & SUNDHEIM 1995] Grishman, R. und Sundheim, B. (1995). *Message Understandig Conference – 6: A Brief History*. In: [MUC-6 1995].

[GRISHMAN 1995] Grishman, R. (1995). *The NYU System for MUC-6 or Where is the Syntax?* In: [MUC-6 1995].

[HAAPALAINEN & MAJORIN 1994] Haapalainen, M. und Majorin, A. (1994). *Gertwol. Ein System zur automatischen Wortformerkennung deutscher Wörter*. Lingsoft Oy, Helsinki.

[HAMP & FELDWEG 1997] Hamp, B. und Feldweg, H. (1997). *GermaNet - a Lexical-Semantic Net for German*. In: *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid, 1997.

[HANISCH ET AL. 2003] Hanisch, D., Fluck, J., Mevissen, H.-T., Zimmer, R. (2003). *Playing Biology's Name Game: Identifying Protein Names in Scientific Texts*. In: *Proceeding of Pacific Symposium on Biocomputing 2003*. S. 403-414.

[HELLFRITZSCH 1995] Hellfritzsch, V. (1995). *Namen der Genossenschaften in der ehemaligen DDR*. In: [EICHLER ET AL. 1995]. S 1611-1613.

[HIRSCHMAN 2003] Hirschman L. (2003). *Using biological resources to bootstrap text mining.* Presentation to the Massachusetts Biotechnology Council Informatics Committee. Nur Online: http://www.e-biosci.org/sept/Hirschman.pdf [12.4.2006].

[HOLZFEIND 1979] Holzfeind, E. (1979). *Die Eigennamen-Analyse und Abgrenzung*. In: *Zur Reform der deutschen Orthographie. Materialien der „Internationalen sprachwissenschaftlichen Tagung zur Reform der deutschen Orthographie".* Wien 1978. Mentrup, W., Pacolt, E., Wiesmann L. Heidelberg. S. 41-70.

[HSU & LIN 2001] Hsu, C. und Lin, C. (2001). *A comparison on methods for multi-class support vector machines*. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

[HUMPHREYS ET AL. 1998] Humphreys, K., Gaizauskas, S., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., Wilks. Y. (1998). *University of Sheffield: Description of the LaSIE-II System as used for MUC-7*. In: [MUC-7 1998].

[JNLPBA-2004] *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004).* Collier, N., Ruch, P., Nazarenko, A. (Hrsg.). Geneva, Switzerland.

[JOACHIMS 1998] Joachims, T. (1998). *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. In: *European Conference on Machine Learning (ECML).*

[JOACHIMS 1999] Joachims, T. (1999). *Making large-Scale SVM Learning Practical*. In: *Advances in Kernel Methods - Support Vector Learning*. Schölkopf, B., Burges, C., Smola, A. (Hrsg.). MIT-Press.

[JOACHIMS 2002] Joachims, T.(2002). *Learning to Classify Text Using Support Vector Machines*. Kluwer International Series in Engineering and Computer Science. Kluwer.

[KALVERKÄMPER 1978] Kalverkämper, H. (1978). *Textlinguistik der Eigennamen*. Klett-Cota. Stuttgart.

[KIM 2004] Kim, J-D., Tsujii, J. (2004). *Word Folding: Taking the Snapshot of Words Instead of the Whole*. In: *The Proceedings of the First International Joint Conference on Natural Language Processing*. S. 61-68.

[KIM ET AL. 2004] Kim, J.-D., Ohta, T., Tsuruoka, Y, Tateisi, Y., Collier, N. (2004). *Introduction to the Bio-Entity Task at JNLPBA*. In: [JNLPBA-2004]. S. 70-76.

[KLEIN ET AL. 2003] Klein, K., Smarr, J., Nguyen H., Manning, C.D. (2003). *Named Entity Recognition with Character-Level Models*. In: [CONLL 2003]. S. 180-183.

[KORNAI & THOMPSON 2005] Kornai, A. und Thompson, B. (2005). *Size doesn't matter*. Unveröffentlichter Entwurf. Nur Online: http://www.kornai.com/Drafts/size.pdf [24.4.2006].

[KRIPKE 1972] Kripke, A.S. (1972). *Naming and Necessity*. In: Harmann, G. und Davidson D. (Hrsg.). *Semantics of Natural Language*. Dordrecht. Boston.

[KRUPKA & HAUSMANN 1998] Krupka, R.G. und Haumann, K. (1998). *IsoQuest, Inc: Description of the NetOwl™ Extractor System Used for MUC-7*. In: [MUC-7 1998].

[KUDOH & MATSUMOTO 2000] Kudoh, T. und Matsumoto, Y. (2000). *Use of Support Vector Learning for Chunk Identification*. In: *Proceedings of the Conference on Natural Language Learning (CoNLL-2000)*. Lissabon. S. 142-144.

[LAFFERTY ET AL. 2001] Lafferty, J., McCallum, A., Pereira, F (2002). *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In: *Proceedings of the International Conference on Machine Learning (ICML-2001)*. Williams, MA.

[LEE ET AL. 2003] Lee, K.-J., Hwang, Y-S., Rim, H.-C. (2003). *Two-phase biomedical NE recognition based on SVMs*. In: *Proceedings of ACL 2003 Workshop on Natural Language Processing in Biomedicine*. Sapporo, Japan.

[LEE ET AL. 2004] Lee, C., Hou, W.-J., Chen, H.-H. (2004). *Annotating Multiple Types of Biomedical Entities: A Single Word Classification Approach*. In: [JNLPBA-2004]. S. 80-83.

[LIN ET AL. 2003] Lin, W., Yangarber, R., Grishman, R. (2003). *Bootstrapped Learning of Semantic Classes from Positive and Negative Examples*. In: *Proceedings of the ICML 2003 Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*. Washington, D.C.

[LÖTSCHER 1995] Lötscher, A. (1995). *Namen von Bildungseinrichtungen*. In: [EICHLER ET AL. 1995]. S. 1606-1611.

[MALOUF 2002] Malouf, R. (2002). *A comparison of algorithms for maximum entropy parameter estimation*. In: [CONLL 2002]. S. 49-55.

[MANDL & WORMSER-HACKER 2005] Mandl, T. und Womser-Hacker, C. (2005). *How do Named Entities Contribute to Retrieval Effectiveness?* In: *Evaluation of Cross-Language Information Retrieval Systems - Proceedings of the CLEF 2004 Workshop*. Peters, C., Clough, P., Gonzalo, J., Kluck, M., Jones, G., Magnini, B. (Hrsgs). Springer, Berlin. S. 649-654.

[MANDL ET AL. 2005] Mandl, T.,Schneider, R., Schnetzler, P., Womser-Hacker, C. (2005). *Evaluierung von Systemen für die Eigennamenerkennung im cross-lingualen Information Retrieval*. In: *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen – Beiträge zur GLDV-Tagung 2005*. Fisseni, B., Schmitz H.-C., Schröder, B., Wagner, P. (Hrsg.). Bonn.

[MANI ET AL. 1996] Mani, I., Macmillan, T.R. , Luperfoy, S., Lusher, E.P., Laskowski, S.J. (1996). *Identifying Unknown Proper Names in Newswire Text*. In: [BOGUAREV & PUSTEJOVSKY 1996]. S. 44-54.

[MARSH & PERZANOWSKI 1998] Marsh E, und Perzanowski D. (1998): *MUC-7 evaluation of IE technology: Overview of results*. In: [MUC-7 1998].

[MAYFIELD ET AL. 2003] Mayfield, J., McNamee, P., Piatko, C. (2003). *Named Entity Recognition using Hundreds of Thousands of Features*. In: [CONLL 2003]. S. 184-187.

[MCCALLUM & LI 2003] McCallum, A. & Li, W. (2003). *Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons*. In: [CONLL 2003]. S. 188-191.

[MCCALLUM 2003] McCallum, A. (2003). *Efficiently Inducing Features of Conditional Random Fields*. In: *Conference on Uncertainty in Artificial Intelligence (UAI)*. Acapulco, Mexico.

[MCDONALD 1996] McDonald, D.D. (1996). *Internal and External Evidence in the Identification and Semantic Categorization of Proper Names*. In: [BOGUAREV & PUSTEJOVSKY 1996]. S. 21-39.

[MEDLINE] Medical Literature Analysis and Retrieval System Online. Öffentlich zugängliche Datenbank des National Center for Biotechnology Information. Online-Zugriff: http://www.ncbi.nlm.nih.gov/entrez/ [12.4.2006].

[MIKHEEV 1997] Mikheev, A. (1997). *Automatic rule induction for unknown-word guessing*. In: *Computational Linguistics, 23(3)*. S. 405–423.

[MIKHEEV ET AL. 1998] Mikheev, A., Groover, C., Moens, M. (1998). *Description of the LTG System Used for MUC-7*. In: [MUC-7 1998].

[MILLER ET AL. 1998] Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., Weischedel, R., the Annotation Group (1989). *Algorithms that Learn to Extract Information – BBN: Description of the Sift System As Used for MUC-7*. In: [MUC-7 1998].

[MUC-6 1995]. *Proceedings of the Sixth Message Understanding Conference*. Morgan Kaufmann. Chinchor, N. (Hrsg.). Columbia, Maryland.1995.
Online: http://acl.ldc.upenn.edu/M/M95/ [12.4.2006].

[MUC-7 1998]. *Proceedings of the Seventh Message Understanding Conference*. Chinchor, N. (Hrsg.). Morgan Kaufmann. Fairfax, Virginia.1998. Online: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html [12.4.2006].

[MUC-APPENDIX 1995] Sixth Message Understanding Conference. *Appendix C: Named Entity Task Definition Version 2.1*. In: [MUC-6 1995]. S. 317-332.

[MÜLLER & KUTAS 1997] Müller, H.M. und Kutas, M. (1997). *Die Verarbeitung von Eigennamen und Gattungsbezeichnungen. Eine elektrophysiologische Studie*. In: Rickheit, G. (Hrsg.). *Studien zur Klinischen Linguistik: Methoden, Modelle, Intervention*. Opladen. Westdeutscher Verlag. S. 147-169.

[NEUMANN & PISKORKSI 2002] Neumann, G. und Piskorksi, J. (2002). *A Shallow Text Processing Core Engine*. In: *Journal of Computational Intelligence,* Volume 18, Number 3. S. 451-476.

[ONTOTEXT 2003] *KIM – Semantic Annotation Platform*. Nur Online verfügbar: http://www.ontotext.com/kim/KIMPlatform.pdf [12.4.2006].

[PALMER & DAY 1997] Palmer, D. D. und Day, D.S. (1997). *A statistical profile of the named entity task.* In: *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP).* Washington, D.C. S. 190-193.

[PASTRA ET AL. 2002] Pastra, K., Maynard, D., Hamza, O., Cunningham H., Wilks, Y. (2002). *How feasible is the reuse of grammars for Named Entity Recognition*? In: *Proceedings of the Language Resources and Evaluation Conference (LREC 2002)*, Las Palmas. S. 1412-1418.

[PEARSON 2001] Pearson, H. (2001). *Biology's name game*. In: *Nature*, 411. S. 631 – 632.

[PISKORKSI & NEUMANN 2000] Piskorski, J. und Neumann, G. (2000). *An intelligent text extraction and navigation system*. In: *Proceedings of 6th International Conference on Computer-Assisted Information Retrieval (RIAO-2000).* Paris.

[PORTER 1980] Porter, M.F. (1980). *An algorithm for suffix stripping*. In: *Program*, 14(3), 1980. S. 130-137.

[QUASTHOFF & BIEMANN 2002] Quasthoff, U. und Biemann, C. (2002). *Named entity learning and verification: EM in large corpora*. In: [CONLL 2002]. S. 8-14.

[QUINLAN 1993] Quinlan, R. J. (1993). *C4.5: Program for Machine Learning*. Morgan Kaufmann Publishers.

[RABINER 1989] Rabiner, L. R. (1989). *A tutorial on Hidden Markov Models and selected applications in speech recognition.* In: *Proceedings of the IEEE,* 77(2), 1989. S. 257-286.

[RAMSHAW & MARCUS 1995] Ramshaw, L.A. und Marcus, M.P. (1995). *Text Chunking using Transformation-based Learning*. In: *Proceeding of the 3rd ACL Workshop on Very Large Corpora at ACL'95*.

[RAU 1991] Rau, L. F. (1991). *Extracting Company Names from Text*. In: Tim Finin (Hrsg.). *Proceedings of the Sixth IEEE Conference on Artificial Intelligence Applications*. IEEE Computer Society Press, Miami Beach, Florida, February 1991. S. 189-194.

[RILOFF & JONES 1999] Riloff, E., & Jones, R. (1999). *Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping*. In: *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*. S. 474-479.

[RILOFF 1993] Riloff, E. (1993). *Automatically Constructing a Dictionary for Information Extraction Tasks*. In: *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93)*. AAAI Press/The MIT Press. S. 811-816.

[RILOFF 1996] Riloff, E. (1996). *An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains*. In: *AI Journal*, Volume 85, August 1996.

[RINALDI ET AL. 2004] Rinaldi, F., Schneider, G., Kaljurand, K., Dowdall, J., Andronis, C., Persidis, A., Konstanti, O. (2004). *Mining Relations in the GENIA corpus*. In: *Proceedings of the second workshop on Data Mining and Text Mining for Bioinformatics*. Pisa, Italy.

[RISTAD 1998] Ristad, E.S. (1998). *Maximum Entropy Modelling Toolkit*, release 1.6 beta. Februar 1998.

[RÖSSLER & MORIK 2005] Rössler, M. und Morik, K. (2005). *Using Unlabeled Texts for Named-Entity Recognition*. In: *Proceedings of the Workshop on Learning with Multiple Views*. 22nd International Conference on Machine Learning. Bonn, Germany.

[RÖSSLER 2004a] Rössler, M. (2004). *Adapting an NER-System for German to the Biomedical Domain*. In: [JNLPBA-2004]. S. 92-95.

[RÖSSLER 2004b] Rössler, M. (2004). *Corpus-based Learning of Lexical Resources for German Named Entity Recognition*. In: *Proceedings of Language Resources and Evaluation Conference (LREC 2004)*. Lisboa. S. 705-708

[RÜPING 2004] Rüping, S. (2004). *A Simple Method for Estimating Conditional Probabilities in SVMs*. In: Abecker, A., Bickel, S., Brefeld, U., Drost, I., Henze, N., Herden, O., Minor, M., Scheffer, T., Stojanovic, L., Weibelzahl, S. (Hrsg.). *LWA 2004 - Lernen - Wissensentdeckung – Adaptivität.* Humboldt-Universität Berlin, 2004.

[SCHMID 1995] Schmid, H. (1995). *Improvements in Part-of-Speech Tagging with an Application to German*. In: *Proceedings of EACL-SIGDAT 1995*. Dublin, Ireland.

[SEARLE 1958] Searle, J.R. (1958). *Proper Names*. In: *Mind* 67. S 166-173.

[SEKINE 1998] Sekine, S. (1997). *NYU system for Japanese NE – MET2*. In: [MUC-7 1998].

[SETTLES 2004] Settles, B. (2004). *Biomedical Named Entity Recognition Using Conditional Random Fields and Novel Feature Sets*. In: [JNLPBA-2004]. S. 104-107.

[SHA & PEREIRA 2003] Sha, F. und Pereira, F. (2003). *Shallow Parsing with Conditional Random Fields.* In: *Proceedings of HLT-NAACL 2003*. Edmonton. S. 134-141.

[SHANNON 1948] Shannon, C. E. (1948). *A mathematical theory of communication*. In: *Bell System Technical Journal*. Volume 27. S. 379-423 und 623-656.

[STEINER 2001] Steiner, I. (2001). *Warum `Named Entities' für die Chunk-Analyse wichtig sind.* In: *Proceedings der GLDV-Frühjahrstagung 2001*. Gießen, S. 245-252.

[STEINER 2002] Steiner, P. (2002). *Das revidierte Münsteraner Tagset / Deutsch (MT/D) – Beschreibung, Anwendung, Beispiele und Problemfälle.* Arbeitsbereich Linguistik, Universität Münster, Version vom 4. Januar 2002.
Online: http://xlex.uni-muenster.de/Portal/MTPD/tagsetDescriptionDE.ps [24.4.2006].

[STRAWSON 1958] Strawson, P.F. (1950). *On Referring*. In: *Mind* 59, S. 320-344. Deutsche Übersetzung in [WOLF 1993], S. 94-126.

[SUNDHEIM 1995] Sundheim, B. (1995). *MUC-6 named entity task definition*, Version 2.1. In: [MUC-6 1995].

[SVM^{Light}] SVM Implementation von Thorsten Joachims. Siehe [JOACHIMS 1999]. Online erhältlich unter: http://svmlight.joachims.org/ [12.4.2006].

[TAKEUCHI & COLLIER 2002] Takeuchi, K. und Collier, N. (2002). *Use of support vector machines in extended named entity recognition*. In: *Proceedings of the Sixth Workshop on Computational Language Learning (CoNLL-2002)*. Taipei, Taiwan. Morgan Kaufmann.

[THELEN & RILOFF 2002] Thelen, M. und Riloff, E. (2002). *A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts*. In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*. Philadelphia.

[THIELEN 1995] Thielen, C. (1995). *An Approach to Proper Name Tagging for German*. In: *From Texts to Tags: Issues in Multilingual Language Analysis - Proceedings of the ACL SIGDAT Workshop*. Dublin. S. 35- 40.

[TJONG KIM SANG & BUCHHOLZ 2000] Tjong Kim Sang, E.F. und Buchholz, S. (2000). *Introduction to the CoNLL-2000 shared task: Chunking*. In: *Proceedings of the Conference on Natural Language Learning (CoNLL-2000)*. S. 127-132.

[TJONG KIM SANG & DE MEULDER 2003] Tjong Kim Sang, E.F und De Meulder F. (2003). *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.* In: [CONLL 2003]. S. 142-147.

[TJONG KIM SANG & VEENSTRA 1999] Tjong Kim Sang, E.F und Veenstra, J. (1999). *Representing Text Chunks*. In: *Proceedings of Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL `99)*. Bergen, Norway. S. 173-179.

[TJONG KIM SANG 2002a] Tjong Kim Sang, E.F. (2002). *Memory-Based Named Entity Recognition*. In: [CONLL 2002]. S. 203-206.

[TJONG KIM SANG 2002b] Tjong Kim Sang, E.F. (2002). *Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition*. In: [CONLL 2002]. S. 155-158.

[TORAL & MUÑOZ 2006] Toral, A., Muñoz, R. (2006). *A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia*. In: *Proceedings of the EACL-2006 Workshop on NEW TEXT - Wikis and blogs and other dynamic text sources*. Trento, Italy.

[VAN RIJSBERGEN 1979] van Rijsbergen, C. (1979). *Information Retrieval*. Butterworths. London.

[VAPNIK 1995] Vapnik, V. N (1995). *The Nature of Statistical Learning Theory*. Springer Verlag. 1995.

[VITERBI 1967] Viterbi, A. J. (1967). *Error Bounds for Convolutional Codes and an Asympotically Optimum Decoding Algorithm*. In: *IEEE Transactions on Information Theory*, IT-13(2). S. 260-269.

[VOLK & CLEMATIDE 2001] Volk, M. und Clematide, S. (2001). *Learn-Filter-Apply-Forget. Mixed Approaches to Named Entity Recognition*. In: *Proceedings of 6th International Workshop on Applications of Natural Language for Information Systems*. Madrid.

[VOLK & SCHNEIDER 1998] Volk, M. und Schneider, G. (1998). *Comparing a statistical and a rule-based tagger for German*. In: Schröder, B., Lenders,W., Hess, W., Portele, T. (Hrsg). *Computers, Linguistics, and Phonetics between Language and Speech.* Proceedings of the 4th Conference on Natural Language Processing - KONVENS-98. Bonn. S. 125-137.

[WACHOLDER ET AL. 1997] Wacholder, N., Ravin, Y., Choi, M. (1997). *Disambiguation of proper names in text*. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*. Washington, D.C. S. 202–208.

[WAKAO ET AL. 1996] Wakao, T., Gaizauskas, R., Wilks, Y. (1996). *Evaluation of an Algorithm for the Recognition and Classification of Proper Names*. In: *Proceedings of the 16th International Conference on Computational Linguistics (COLING96)*. Copenhagen. S. 418-423.

[WEISCHEDEL ET AL. 1993] Weischedel, R., Meteer, M., Schwartz, R., Ramshaw, L., Palmucci, J. (1993). *Coping with Ambiguity and Unknown Words through Probabilistic Methods*. In: *Computational Linguistics*, 19(2). S. 359-382.

[WIKIPEDIA] Wikipedia – Die freie Enzyklopädie. http://www.wikipedia.de [24.04.2006].

[WOLF 1993] *Eigennamen – Dokumentation einer Kontroverse*. Wolf, U. (Hrsg.). Frankfurt am Main. Suhrkamp.

[ZHOU & SU 2002] Zhou, G. und Su, J. (2002). *Named Entity Recognition using an HMM-based Chunk Tagger*. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).* Philadelphia. S. 473-480.

[ZHOU & SU 2003] Zhou, G. und Su, J. (2003). *Integrating Various Features in Hidden Markov Model using Constraint Relaxation Algorithm for Recognition of Named Entities without Gazetteers*. In: *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering.* Peking. S. 732-739.

[ZHOU & SU 2004] Zhou, G. und Su, J. (2004). *Exploring Deep Knowledge Resources in Biomedical Name Recognition*. In: [JNLPBA-2004]. S. 96-99.